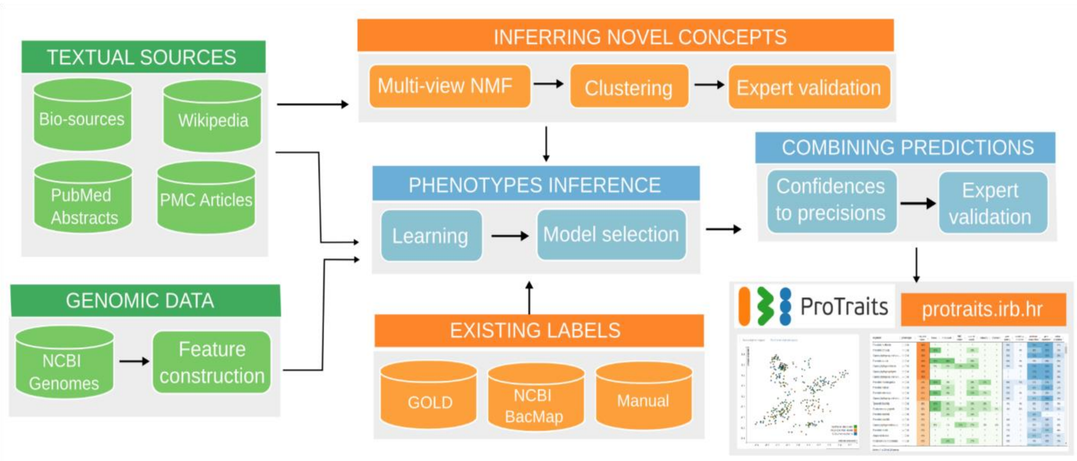
Učenje reprezentacije za podatke iz jednog i više pogleda

Maria Brbić
Mentor: Dr. sc. Ivica Kopriva
Fakultet elektrotehnike i računarstva, Sveučilište u Zagrebu
Institut Ruđer Bošković, Zavod za elektroniku

Uvod. Uspjeh algoritama strojnog učenja uvelike ovisi o reprezentaciji podataka. Glavni cilj doktorskog istraživanja je razvoj algoritama za učenje reprezentacija podataka iz jednog i više pogleda. Visoko-dimenzionalni podatci često leže u nisko-dimenzionalnim potprostorima. Uz smanjivanje računalnih resursa, pronalazak nisko-dimenzionalne reprezentacije je od iznimne je važnosti za smanjivanje šuma prisutnog u visoko-dimenzionalnim podatcima i poboljšavanje točnosti algoritama strojnog učenja. Također, mnogi realni podaci su opisani s heterogenim reprezentacijama ili pogledima. U slučaju grupiranja podataka iz više pogleda, tipična je pretpostavka da je pripadnost primjera grupama dijeljena između pogleda. Iako je svaki pogled zasebno dovoljan za grupiranje, kombiniranjem pogleda moguće je postići veću točnost.

Metodologija za automatsko učenje fenotipova iz više pogleda



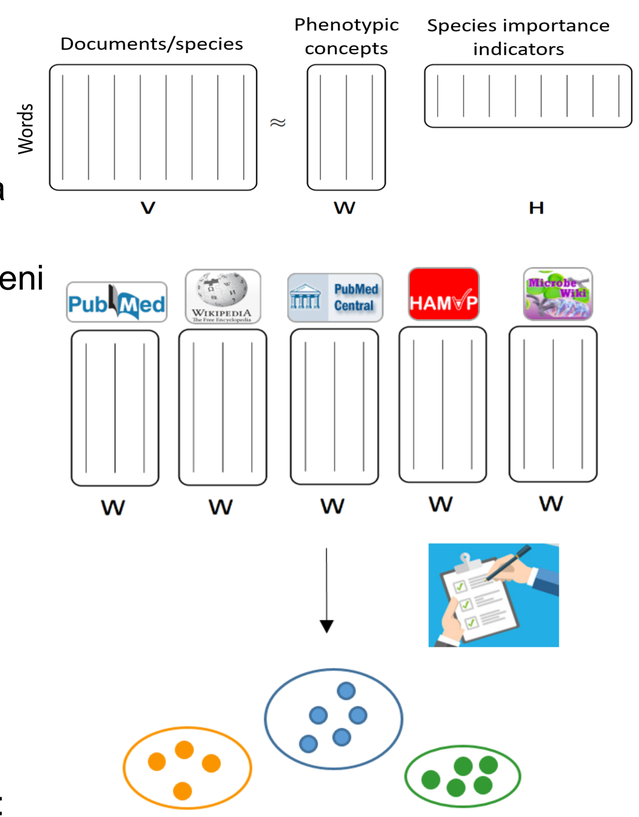
Glavne komponente ProTraits sustava:

i. Komponenta za otkrivanje novih fenotipskih koncepata iz više tekstnih izvora temeljena na metodi nenegativne matrične faktorizacije
ii. Komponenta za predviđanje fenotipova iz više tekstnih izvora i genomskih reprezentacija koristeći algoritme nadziranog strojnog učenja
iii. Komponenta za kombinaciju predikcija temeljena na kasnoj intergraciji

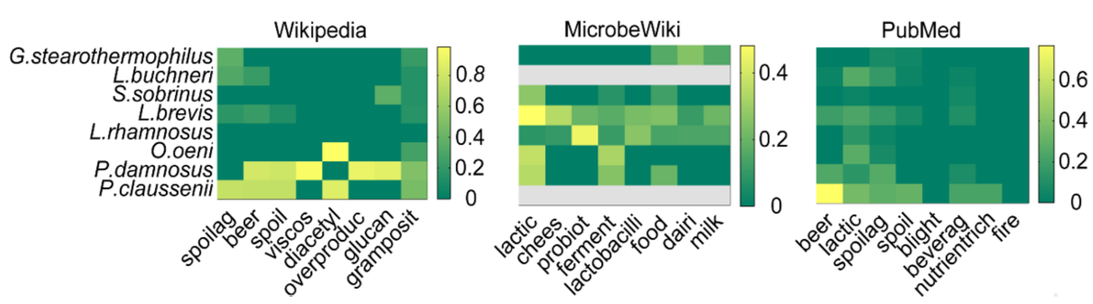
Učenje novih fenotipskih koncepata. Koristeći ne-negativnu matričnu faktorizaciju na podacima iz više tekstnih izvora otkrili smo više od 100 novih fenotipskih koncepata.

Pristup:

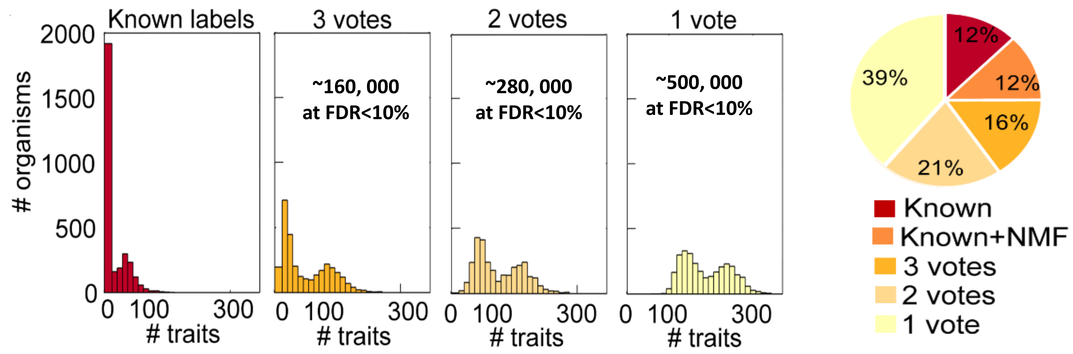
i. Primijeniti NMF na pojedinačnim pogledima
ii. Grupirati NMF faktore zahtijevajući da je otkriveni koncept konzistentno prisutan u barem tri pogleda kako bi se maksimizirala različitost koncepata
iii. Ponoviti za različiti broj faktora i s različitim inicijalizacijama
iv. Provjera ljudskog stručnjaka



Primjer otkrivenog fenotipa: uzrokuje li mikrob kvarenje piva?



Rezultati



Ukupni broj organizama pokrivenih s fenotipovima koristeći ProTraits predikcije zahtijevajući slaganje barem tri, dva ili jedan tekstni ili genomski izvor.

Algoritmi učenja reprezentacije iz više pogleda (MLRSSC)

Ključan korak u primjeni algoritama grupiranja u potprostorima je konstrukcija robusne matrice susjedstva.

Formulacija problema: Uz zadane skupove podataka $\mathbf{X}=\{\mathbf{X}^{(1)},\ldots,\mathbf{X}^{(v)}\}$, pri čemu je svaki $\mathbf{X}^{(i)}$ opisan sa svojim skupom značajki, cilj je pronaći reprezentacijsku matricu $\mathbf{C}$ zajedničku kroz sve poglede.

Ciljna funkcija za učenje rijetke reprezentacijske matrice niskog ranga za grupiranje u potprostorima podataka iz više pogleda:

i. Uz poticanje sličnosti između parova reprezentacijskih matrica (PMLRSSC):

$$\min_{\mathbf{C}^{(1)},\mathbf{C}^{(2)},\ldots,\mathbf{C}^{(n_v)}} \sum_{v=1}^{n_v} \left(\beta_1 \|\mathbf{C}^{(v)}\|_* + \beta_2 \|\mathbf{C}^{(v)}\|_1\right) + \sum_{1\le v,w\le n_v, v\neq w} \lambda^{(v)} \|\mathbf{C}^{(v)} - \mathbf{C}^{(w)}\|_F^2$$
$$s.t. \quad \mathbf{X}^{(v)} = \mathbf{X}^{(v)}\mathbf{C}^{(v)}, \quad diag(\mathbf{C}^{(v)}) = \mathbf{0}, \quad v = 1,\ldots n_v,$$

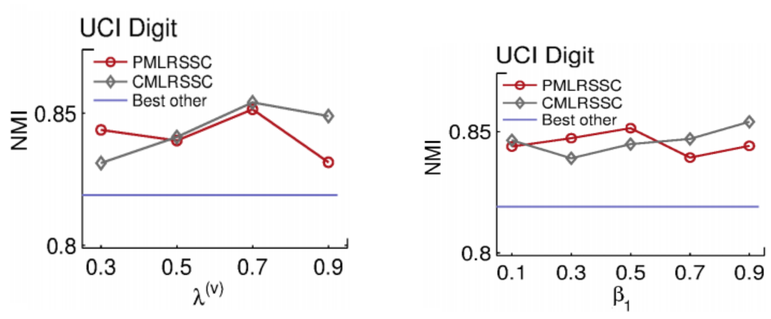
ii. Uz poticanje reprezentacijskih matrica prema centroidu (CMLRSSC):

$$\min_{\mathbf{C}^{(1)},\mathbf{C}^{(2)},\ldots,\mathbf{C}^{(n_v)}} \sum_{v=1}^{n_v} \left(\beta_1 \|\mathbf{C}^{(v)}\|_* + \beta_2 \|\mathbf{C}^{(v)}\|_1 + \lambda^{(v)} \|\mathbf{C}^{(v)} - \mathbf{C}^{*}\|_F^2\right)$$
$$s.t. \quad \mathbf{X}^{(v)} = \mathbf{X}^{(v)}\mathbf{C}^{(v)}, \quad diag(\mathbf{C}^{(v)}) = \mathbf{0}, \quad v = 1,\ldots n_v,$$

Rješenje optimizacijskih problema temelji se na metodi ADMM (engl. *Alternating Direction Method of Multipliers*).
MLRSSC algoritmi uče matricu susjedstva koja modelira linearnu strukturu potprostora. S ciljem modeliranja nelinearna strukture predložene su ekstenzije algoritama → kernel MLRSSC algoritmi.

Rezultati

	3-sources	Reuters	UCI Digit	Prokaryotic	Synthetic
Best Single View					
LRSSC	0.569	0.333	0.702	0.579	0.624
Feature Concat LRSSC	0.579	0.347	0.698	0.584	0.682
Co-reg Pairwise	0.463	0.371	0.694	0.468	0.660
Co-reg Centroid	0.505	0.362	0.754	0.459	0.646
RMSC	0.477	0.361	0.742	0.447	0.715
CSMSC	0.482	0.365	0.775	0.462	0.730
Pairwise MLRSSC	**0.659**	0.428	0.830	**0.591**	0.689
Centroid MLRSSC	**0.654**	**0.432**	**0.835**	0.574	0.690
Pairwise KLMRSSC	0.541	**0.429**	0.827	**0.591**	**0.742**
Centroid KMLRSSC	0.556	0.426	**0.840**	**0.582**	**0.743**



Osjetljivost parametara na UCI Digit skupu podataka.

Najvažnije publikacije:

Brbić, M.; Kopriva, I. Multi-view Low-rank Sparse Subspace Clustering. *Pattern Recognition*. 73 (2018)
Brbić, M.; Piškorec, M.; Vidulin, V.; Kriško, A.; Šmuc, T.; Supek, F. The Landscape of Microbial Phenotypic Traits and Associated Genes. *Nucleic Acids Research*. 44 (2016)

Zahvale. Ovaj doktorat je nastao u okviru EU FP7 projekta MAESTRA, te HrZZ projekata "Strukturne dekompozicije empirijskih podataka za računalno potpomognutu dijagnostiku bolesti" i "Algoritmi strojnog učenja za otkrivanje znanja u složenim strukturama podataka".

Dan doktorata Fakulteta elektrotehnike i računarstva, 10. svibnja 2018.