Jumping across contexts using compressive data fusion

Marinka Zitnik
Stanford University
Department of Computer Science
Heterogeneous associations

Molecular functions

Genes

MTH1

Diseases

Breast neoplasms

Symptoms

Swollen lymph nodes

Environmental exposure

Ionizing radiation
Heterogeneous associations

- Molecular functions
- Diseases
- Symptoms
- Genes
- Environmental exposure
Heterogeneous associations

Set of associations between objects A and objects B
Heterogeneous associations

Set of associations between objects A and objects B
Heterogeneous associations

- Gene Ontology terms
- Chemicals
- Pathways
- Diseases
- Genes
- Molecular signatures
- Exposure events
- Disease symptoms

Set of associations between objects A and objects B
Heterogeneous associations

Set of associations between objects A and objects B
Different ways of relating objects carry different semantic meanings and result in different predictions!

Mixing, discarding or ignoring semantics can impede model performance and its explanatory capability!
Modeling heterogeneous associations

Understanding heterogeneous associations by connecting biomedical contexts

**Prediction:** Can we estimate whether a gene and a disease are likely to be associated? Can we find a set of genes that, taken together, appear most significant to another set of genes?

**Understanding:** Can we understand different similarities that are implied by different semantics? Can we use this knowledge to help user choose/combine the semantics?
Problem definition

• **Size-k module**: A set of $k$ objects with a shared property

• **Candidates**: A pool of objects from which a module is identified

• **Pivots**: Objects considered to assess significance of the module

• **Goal**: Find a size-$k$ module of candidates that is maximally significant with respect to the given pivots
Model for module detection

Let $V$ denote the candidates and $f : 2^V \rightarrow \mathbb{R}_+$ be a module scoring function.

Function $f(X)$ measures the "goodness" of $X$ in the sense of $X$ being significant relative to the pivots.

This leads to the optimization problem:

$$\text{find } X^* \text{ such that } f(X^*) = \max_{Y \subset V, |Y|=k} f(Y)$$

Bad news: Solving the problem exactly is NP-complete (Feige, 1998)
The nature of the goodness function

\[ V = \{ g_1, g_2, g_3, g_4, g_5, g_6, g_7, g_8 \} \]
The nature of the goodness function

\[ V = \{ g_1, g_2, g_3, g_4, g_5, g_6, g_7, g_8 \} \]

\[ f(g_5, g_6, g_7, g_8) = \text{goodness of considering genes together} \]
The nature of the goodness function

$$V = \{ g_1, g_2, g_3, g_4, g_5, g_6, g_7, g_8 \}$$

$$f(\{g_5, g_6, g_7, g_8\}) = \text{goodness of treating genes together}$$

$$S \subseteq T$$

$$f(S \cup g_6) - f(S) \geq f(T \cup g_6) - f(T)$$

**Diminishing marginal gains**
The nature of the goodness function

$S \subseteq T$

$$f(S \cup g_6) - f(S) \geq f(T \cup g_6) - f(T)$$

- Significance of $g_6$ is discounted based on how much of its representation exists in the current module.
- The value of $g_6$ diminishes as the module grows:
  - $g_6$ is added to the module when its marginal gain is the largest.
  - If $g_6$ is added to the module later, its marginal gain is smaller.
That’s submodular optimization

- The goodness function $f$ is a submodular set function that is also nonnegative and monotone non-decreasing
- We inherit 40+ years of research!
- Good news:
  When $f$ is a monotone, submodular and nonnegative function on $V$, then the greedy algorithm has the worst case guarantee:

  $$f(X) \geq (1 - 1/e)f(X^*)$$

  where $X$ is the greedy solution and $X^*$ is the optimal solution
Size-$k$ maximally significant module

- Estimate the semantic of interest using a collective latent model
- Start with an empty module, pull in one candidate at a time
- Repeat $k$ times:
  - Compute concentration significance for all the candidates
  - Pull candidate $s$ with the highest score into the module
  - Expand pivot set with candidate $s$
Size-$k$ maximally significant module

- Estimate the semantic of interest using a collective latent model
- Start with an empty module, pull in one candidate at a time
- Repeat $k$ times:
  - Compute concentration significance for all the candidates
  - Pull candidate $s$ with the highest score into the module
  - Expand pivot set with candidate $s$
Size-\(k\) maximally significant module

- Estimate the semantic of interest using a collective latent model.
- Start with an empty module, pull in one candidate at a time.
- Repeat \(k\) times:
  - Compute concentration significance for all the candidates.
  - Pull candidate with the highest score into the module.
  - Expand pivot set with candidate.
- Estimate the semantic via collective matrix co-factorization.
  (Zitnik et al. IEEE TPAMI 2015)

\[
f (g_5, g_6, g_7, g_8) = \text{relevance + diversity or coverage}
\]

concentration significance
Predicting gene-disease associations

- **Data**
  - 16 datasets, 13 object types
  - Prevent contamination with known gene-disease associations

- **Setup**
  - For each disease in the corpus, do leave-one-gene-out cross-validation, use the remaining genes to fit the model
  - Report AUROC, AUPRC
Considered semantics

Combined semantics
Predicted gene-disease associations

Combined semantics

AUPRC

0.2  0.4  0.6  0.8  1.0
Predicted gene-disease associations

Performance varies substantially across semantics
It provides insights into the utility of semantics
The combined predictor is the most accurate
Detecting disease modules

- **Data**
  - Similar as for the gene-disease association prediction task

- **Setup**
  - For each disease module in the corpus, randomly remove 25%, 50%, or 75% of the disease genes
  - Remaining genes are pivots
  - Detect size-$k$ module, $k$ is the size of full module
  - Report recall
Detecting disease modules

Average recall

Monogenic disease
Cognitive disorder
Endocrine system disease
Orofacial cleft
Immune system disease
Viral infectious disease
Nervous system disease
Unknown
Cardiovascular system disease
Urinary system disease
Integumentary system disease
Acquired metabolic disease
Gastrointestinal system disease
Musculoskeletal system disease
Respiratory system disease
Benign neoplasm
Cancer
Parasitic infectious disease
Psoriatic arthritis
Congenital nervous system abnormality
Developmental disorder of mental health
Inherited metabolic disorder
Sleep disorder
Reproductive system disease
Substance-related disorder

Semantics
• Combined analysis of all semantics performs best

Detecting disease modules

Average recall

0.4
0.3
0.2
0.1

-0.1

Semantics

- Monogenic disease
- Cognitive disorder
- Endocrine system disease
- Orofacial cleft
- Immune system disease
- Viral infectious disease
- Nervous system disease
- Unknown
- Cardiovascular system disease
- Urinary system disease
- Integumentary system disease
- Acquired metabolic disease
- Gastrointestinal system disease
- Musculoskeletal system disease
- Respiratory system disease
- Benign neoplasm
- Cancer
- Parasitic infectious disease
- Psoriatic arthritis
- Congenital nervous system abnormality
- Developmental disorder of mental health
- Inherited metabolic disorder
- Sleep disorder
- Reproductive system disease
- Substance-related disorder
• Combined analysis of all semantics performs best

• Related diseases show similar recall pattern across different semantics
Detecting disease modules

- Combined analysis of all semantics performs best
- Related diseases show similar recall pattern across different semantics
- Highest true positive rates are achieved in the early iterations of the algorithm
- The order in which the candidates are pulled into the module reflects their relevance to the pivots
Recap

- Beyond disease module detection: drug-target prediction, disease-exposure prediction, gene prioritization
- *Collective latent factor model* admits studying different ways of relating objects to each other
- *Submodularity* admits theoretical guarantees
Medusa

An approach to detect size-\(k\) modules of objects that, taken together, appear most significant to another set of objects

github.com/marinkaz/medusa

Travel funding to ISMB 2016 was generously provided by HiTSeq.