Redundancy-aware unsupervised ranking based on game theory - application to gene enrichment analysis

Chiara Balestra
Technische Universität Dortmund
Dortmund, Germany

Emmanuel Müller
Technische Universität Dortmund
Dortmund, Germany

Carlo Maj
IGSB, University Hospital Bonn
Bonn, Germany

Andreas Mayr
IMBIE, University Hospital Bonn
Bonn, Germany

Due the huge amount of data in bioinformatics application, supervised and unsupervised feature selection became essential tools. These methods are often based on individual features subsets' scores, where the crucial question is how to quantify the importance of single features. Several feature scores appeared in the literature but they often ignore the correlations among features selves. Game theory and Shapley values play an important role in recent machine learning literature. Shapley values offer a direct way to rank features based on their importance in supervised contexts; this explains their success and diffusion in various fields.

In computational biology, reducing the overlap among pathways within gene sets gained interest in the recent years; smaller gene-sets with decreased intersection among pathways would be computationally more treatable and the understanding of the pathways selves by specialist could take advantages of this.

Game Theory and Shapley values

Cooperative game theory (CGT) allows to fairly allocate resources among players. In the recent literature, CGT found application in computer science community in feature selection [3], [10] and Shapley values have been adapted for bioinformatics applications in order to test genotype/phenotype association or for gene sets prioritization analysis [11]. One of the advantages of using cooperative game theory for feature selection is the flexible and non-demanding definition of the *value function* which quantifies the resources to be fairly allocated among players. Shapley values' exact computation on the other hand requires 2^N evaluations of the value function where N is the number of players; this makes its application unfeasible as soon as the number of players increases. Microarray games [8] reduce the computational challenges in the computation of exact Shapley values to polynomial time under the assumption that it is possible to express the game using binary relationships.

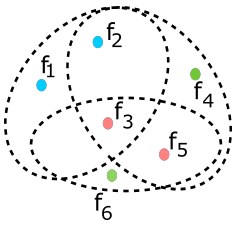

Fig. 1. Each subset of features f_is is considered to compute Shapley values. Correlated features are color-coded.

Application to gene sets and pathways

We extended *microarray games* to pathways and gene sets. Pathways are sets of genes corresponding to functionally related interacting proteins genes while collections of pathways based on prior biological knowledge are denoted as *gene sets* [7]. The pathways in the gene sets can arbitrarily overlap and their sizes spread between tens and thousand of genes. Some recent directions of research try to limit the overlap among pathways while keeping a high coverage of the genes included in the gene sets. The proposed methods include visualization tools of redundancy among pathways, merging pathways based on similarity and integrating full pathways sets into a non-redundant single and unified pathway [1], [4], [12]. However, they all concentrate on modifying the pathways selves rather than reducing the amount of pathways in the gene set through a thoughtful selection.

Methods

We develop a new method based on cooperative game theory and Shapley Values to assign importance scores to pathways in a gene set $\mathcal{F}$. We introduce a set based quantification of resources, where each player is a pathways whose elements are the genes and the set of players is the gene set. The aim is to rank the pathways in $\mathcal{F}$ based on the distribution of the genes in them and the overlapping among the different

pathways. More generally, the obtained rankings allow for a fair evaluation of the *importance* of sets.

Shapley values assign to each pathway an importance score based on the size of the pathway and the distribution among the others pathways of its genes. The first challenge we have to address is the appearance of a correlation among the position in the rank of the pathways and their size. Shapley values tend to assign higher importance scores to high dimensional sets while do not consider the possible shared genes within the pathways. We compute the Shapley values adapting the microarray games to our scope. The binary relationship in the context of pathways and genes is the membership of the genes to the single pathways. As we mentioned, the computation of Shapley values reduces to polynomial run-time allowing for acceptable run-times in the context of gene sets.

The overlap among pathways is a major challenge in this context; we develop a way to address the problem of overlapping pathways while not compromising with the coverage of the genes in the gene set. We integrate in the Shapley values a measure of overlapping among sets: In particular, we consider the *jaccard index*, a well established score to quantify the overlapping among sets A and B; the *Jaccard index* [6] $j(A, B)$ is defined as the ratio among the size of intersection $A \cap B$ and the size of the union $A \cup B$. Punished Shapley values with the jaccard score are then used to rank pathways in the gene sets.

GOALS

Coverage of the gene set: Being $\mathcal{F}$ a gene set and $\{P_i\}_{i=1,\dots,N}$ the pathways in $\mathcal{F}$, we denote with G the genes which are contained in at least one P_i. The *coverage of* G $c_G(\mathcal{S})$ is the percentage of elements $g \in G$ that are included at least in one set when selecting a limited amount of pathways $\mathcal{S} \subseteq \mathcal{P}(\mathcal{F})$.

Decrease overlapping among pathways: Averaging the Jaccard indices of any couple of sets contained in $\mathcal{S}$, we define the *Jaccard rate* $\text{jac}(\mathcal{S})$ of a family of pathways $\mathcal{S}$, i.e.,

$$\text{jac}(\mathcal{S}) = \frac{1}{m(m-1)} \sum_{P_i, P_j \in \mathcal{S}, P_i \neq P} j(P_i, P_j)$$

where $m = |\mathcal{S}|$ and represents the average Jaccard index among any two pathways in $\mathcal{S}$. The Jaccard rate of a family of pathways $\mathcal{S}$ is always a non-negative real number between 0 and 1; it can not reach the upper limit 1 but, it is worth to notice that $\text{jac}(\mathcal{S}) = 0$ if and only if any couple of sets in $\mathcal{S}$ do not intersect. In order to minimize the redundancy in a subset of pathways $\mathcal{S}$, it is necessary to select sets whose intersection is as small as possible.

Study of the impact to GSEA: Gene Set Enrichment Analysis (GSEA) refers to a variety of methods trying to assess the enrichment of genes in different pathways concerning a phenotype with the aim of identifying biological mechanisms potentially associated with a phenotype. Corrections for multiple testing [5] are necessary when testing for significance and potentially lead to a loss of statistical significant pathways [2], [9]. Another more basic approach to avoid the loss of statistical power is to reduce the number of tests to perform. Limiting the number of tested pathways within a gene set w.r.t. to a specific phenotype could lead to a bias while a potential solution might be the incorporation of *unsupervised* approaches to reduce the dimension of the gene set before even considering a specific phenotype. If the reduction of the number of tests needed to be performed is independent of the phenotype and preserves the maximum amount of information contained in the gene set, the typically inflated type-I error due to pre-screening is avoided.

RESULTS

We conducted several experiments and evaluated them with respect to the three mentioned goals. The pathways selected using our pathways ranking show good performances w.r.t. the coverage of the entire gene sets ($\approx 80\%$ of the genes are retained with only the 20% of the pathways). Moreover, the approach is retaining much lower redundancy as expected; the integration of jaccard rate in the Shapley values allowed for a fair ranking with low redundancy. Finally, with respect to an increase statistical power, we are able to show that when considering a lower amount of pathways, we do not assist to a decrease in the number of significant pathways found.

REFERENCES

[1] F. BELINKY, N. NATIV, G. STELZER, S. ZIMMERMAN, T. INY STEIN, M. SAFRAN, AND D. LANCET, *PathCards: multi-source consolidation of human biological pathways*, Database, 2015 (2015).

[2] Y. BENJAMINI AND Y. HOCHBERG, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*, Journal of the Royal Statistical Society. Series B (Methodological), 57 (1995), pp. 289–300. Publisher: [Royal Statistical Society, Wiley].

[3] S. COHEN, G. DROR, AND E. RUPPIN, *Feature selection via coalitional game theory*, Neural Computation, 19 (2007), pp. 1939–1961.

[4] M. S. DODERER, Z. ANGUIANO, U. SURESH, R. DASHNAMOORTHY, A. J. BISHOP, AND Y. CHEN, *Pathway Distiller - multisource biological pathway consolidation*, BMC Genomics, 13 (2012), p. S18.

[5] S. DUDOIT AND M. LAAN, *Multiple Testing Procedures With Applications to Genomics*, Jan. 2008. Journal Abbreviation: Multiple Testing Procedures with Applications to Genomics Publication Title: Multiple Testing Procedures with Applications to Genomics.

[6] P. JACCARD, *Etude de la distribution florale dans une portion des Alpes et du Jura*, Bulletin de la Societe Vaudoise des Sciences Naturelles, 37 (1901), pp. 547–579.

[7] A. LIBERZON, C. BIRGER, H. THORVALDSDÓTTIR, M. GHANDI, J. P. MESIROV, AND P. TAMAYO, *The Molecular Signatures Database (MSigDB) hallmark gene set collection*, Cell Systems, 1 (2015), pp. 417–425.

[8] S. MORETTI, F. PATRONE, AND S. BONASSI, *The class of microarray games and the relevance index for genes*, TOP, 15 (2007), pp. 256–280.

[9] S. NAKAGAWA, *A farewell to Bonferroni: The problems of low statistical power and publication bias*, Behavioral Ecology, 15 (2004), pp. 1044–1045. Place: United Kingdom Publisher: Oxford University Press.

[10] K. PFANNSCHMIDT, E. HÜLLERMEIER, S. HELD, AND R. NEIGER, *Evaluating Tests in Medical Diagnosis: Combining Machine Learning with Game-Theoretical Concepts*, Information Processing and Management of Uncertainty in Knowledge-Based Systems, 610 (2016), pp. 450–461.

[11] M. SUN, S. MORETTI, K. PASKOV, N. STOCKHAM, M. VARMA, B. CHRISMAN, P. WASHINGTON, J.-Y. JUNG, AND D. WALL, *Game theoretic centrality: a novel approach to prioritize disease candidate genes by combining biological networks with the Shapley value*, BMC Bioinformatics, 21 (2020).

[12] M. P. VAN IERSEL, T. KELDER, A. R. PICO, K. HANSPERS, S. COORT, B. R. CONKLIN, AND C. EVELO, *Presenting and exploring biological pathways with PathVisio*, BMC Bioinformatics, 9 (2008), p. 399.