Cancer Immunotherapy: Challenge 3

Dariusz Brzezinski[1] Wojciech Kotlowski[1]

[1]Institute of Computing Science, Poznan University of Technology

Abstract

In this write-up, we propose a metric for assessing the effectiveness of a perturbation. The metric, called Kappa-TVD-LCB (κ_{TL}), is a relative performance measure that takes into account: 1) the target cell state proportions, 2) reference (unperturbed) cell state proportions, and 3) the confidence of the estimation (cell count). The κ_{TL} metric will be presented gradually, with justifications for each step. Moreover, we will comment on using cell state decision class boundaries and present an interactive visualization method tailored to analyzing and selecting the best perturbations.

1 TVD as an absolute performance measure

The first ingredient for our metric is a function assessing the distance between two cell state proportions. For this purpose, we propose to use *total variation distance* (TVD) defined as half of the $l1$-loss:

$$TVD = \frac{1}{2}\|Q_1 - Q_2\|_1 = \frac{1}{2}\sum_{\omega\in\Omega} |Q_1(\{\omega\}) - Q_2(\{\omega\})| \tag{1}$$

where Q_1 and Q_2 are cell state proportions. We propose TVD for comparing two proportions because: it has values in the range [0,1], it is a metric, and it is the optimal transportation cost when the cost function is $c(x, y) = \mathbf{1}_{x\neq y}$[1]. In the following section, we will use TVD to construct a relative measure.

2 Kappa statistic for TVD as a relative performance measure

The Kappa statistic [2] is a popular measure for benchmarking inter-rater agreement and assessing classification accuracy in class imbalance scenarios. The Kappa statistic κ is defined as:

$$\kappa = \frac{p - p_{ran}}{1 - p_{ran}} \tag{2}$$

where p is the accuracy of the classifier under consideration and p_{ran} is the accuracy of the random/baseline classifier. If the predictions of the classifier are perfectly correct then $\kappa = 1$. If its predictions coincide with the correct ones as often as by chance, then $\kappa = 0$. Note that κ can theoretically be negative; this may happen if the classifier predicts worse than chance.

Depending on the context, the baseline classifier can be a majority stub, the last correct prediction, or any other naive method [3]. Given Q as the desired cell state proportion, Q_i as the cell state proportion obtained from knocking out gene i, and Q_0 as the baseline (e.g., unperturbed) proportion, we propose to modify the Kappa statistic to use $1 - TVD(Q, Q_i)$ as p and $1 - TVD(Q, Q_0)$ as p_{ran}:

$$\kappa_T = \frac{(1 - TVD(Q, Q_i)) - (1 - TVD(Q, Q_0))}{1 - (1 - TVD(Q, Q_0))} = \frac{TVD(Q, Q_0) - TVD(Q, Q_i)}{TVD(Q, Q_0)} = 1 - \frac{TVD(Q, Q_i)}{TVD(Q, Q_0)} \tag{3}$$

The proposed κ_T measure is equal to 1 when the desired and knockout cell state proportions match, 0 when the gene knockout is as good as an unperturbed cell state proportion, and negative when the knockout proportion is worse than the unperturbed proportion. Therefore, κ_T is easily interpretable regardless of the desired cell state proportion Q_i and makes it easy to spot knockouts better than unperturbed cell state proportions.

3 Lower confidence bound (LCB) as an uncertainty modification

The κ_T measure takes into account the desired cell state proportion Q, the knockout proportion Q_i, and the unperturbed/baseline proportion Q_0. To take into account the number of cells going into proportion Q_i, we propose to calculate the lower confidence bound of κ_T. The idea is inspired by the UCB1 k-armed bandits algorithm [4] where the estimation of the mean for a given arm is modified by adding an upper confidence bound (UCB) to promote (optimistic) exploration of multiple arms. Here, we want to promote more reliable estimates, and therefore we will subtract from κ_T to obtain a lower confidence bound (LCB).

To calculate LCB we will use the Hoeffding bound.[1] The Hoeffding bound states that with probability 1 the true mean of a random variable of range R does not differ from the estimated mean, after N independent observations, by more than $\epsilon = \sqrt{\frac{R^2 ln(1/\delta)}{2N}}$ [6]. With κ_T having its range defined by $1/TVD(Q, Q_0)$, δ being the allowed estimation error (typically 0.05), and N being the samples size (number of cells), we can define our main proposal the Kappa-TVD-LCB measure (κ_{TL}) as:

$$\kappa_{TL} = \kappa_T - \sqrt{\frac{(1/TVD(Q, Q_0))^2 ln(1/\delta)}{2N}} \tag{4}$$

Let us present of few examples showing the differences between κ_T and κ_{TL}. Let us assume the desired cell state proportion is $Q = (0.95, 0, 0, 0.05, 0)$, the evaluated proportion is $Q_i = (0.37, 0.13, 0.28, 0.20, 0.02)$, and the unperturbed proportion is $Q_0 = (0.0675, 0.2097, 0.3134, 0.3921, 0.0173)$. For the abovementioned input, $\kappa_T(Q, Q_i, Q_0) = 0.343$ showing that Q_i is much better (closer to the desired proportion) than the baseline (unperturbed proportion). If we use the lower confidence bound assuming an allowed error $\delta = 0.05$ and the number of cells $N = 200$, we get $\kappa_{TL}(Q, Q_i, Q_0, 0.05, 200) = 0.245$. If the sample size were smaller ($N = 20$), we get $\kappa_{TL}(Q, Q_i, Q_0, 0.05, 20) = 0.033$. As can be seen, the sample size plays an important role in κ_{TL} and can help assess the quality of the cell state proportion estimate. The implementations of TVD, κ_T, and κ_{TL} can be found in a Jupyter notebook submitted together with this document.

4 Comments on decision boundaries and ideal cell state proportions

We would like to advise against using cell state decision boundaries as part of an evaluation metric. By using the decision boundaries of one machine learning model to evaluate another model, one runs into the risk of biasing the evaluation towards models that are more in line with the model used to create the boundaries. For example, if one uses k-NN to determine cell state proportions and the boundaries of that model are used during evaluation, this may potentially promote distance-based models that predict cell state proportions. Moreover, decision boundaries of cell state proportions may change when more data arrive, especially if some non-deterministic method such as UMAP is part of the process. Finally, the boundaries will also change should the number of analyzed cell states change from five to any other number. Considering all of the above problems with decision boundaries, we oppose to their use in evaluation measures.

As for suggesting ideal state proportions for fighting cancer, as computer scientists, we are ill-equipped for providing detailed guidelines. However, we believe that the most reliable source for such estimations lies in patient data. Ideally, through whole-genome sequencing or RNA sequencing of responders and non-responders to different therapies, one may find genomic variants that increase the chances of successfully fighting tumors. Such variants and their accompanying T-cell properties might not be a direct way to find a cell state proportion for primary treatment, but could be a successful way of finding adjuvant therapies and a milestone towards understanding T-cells even more.

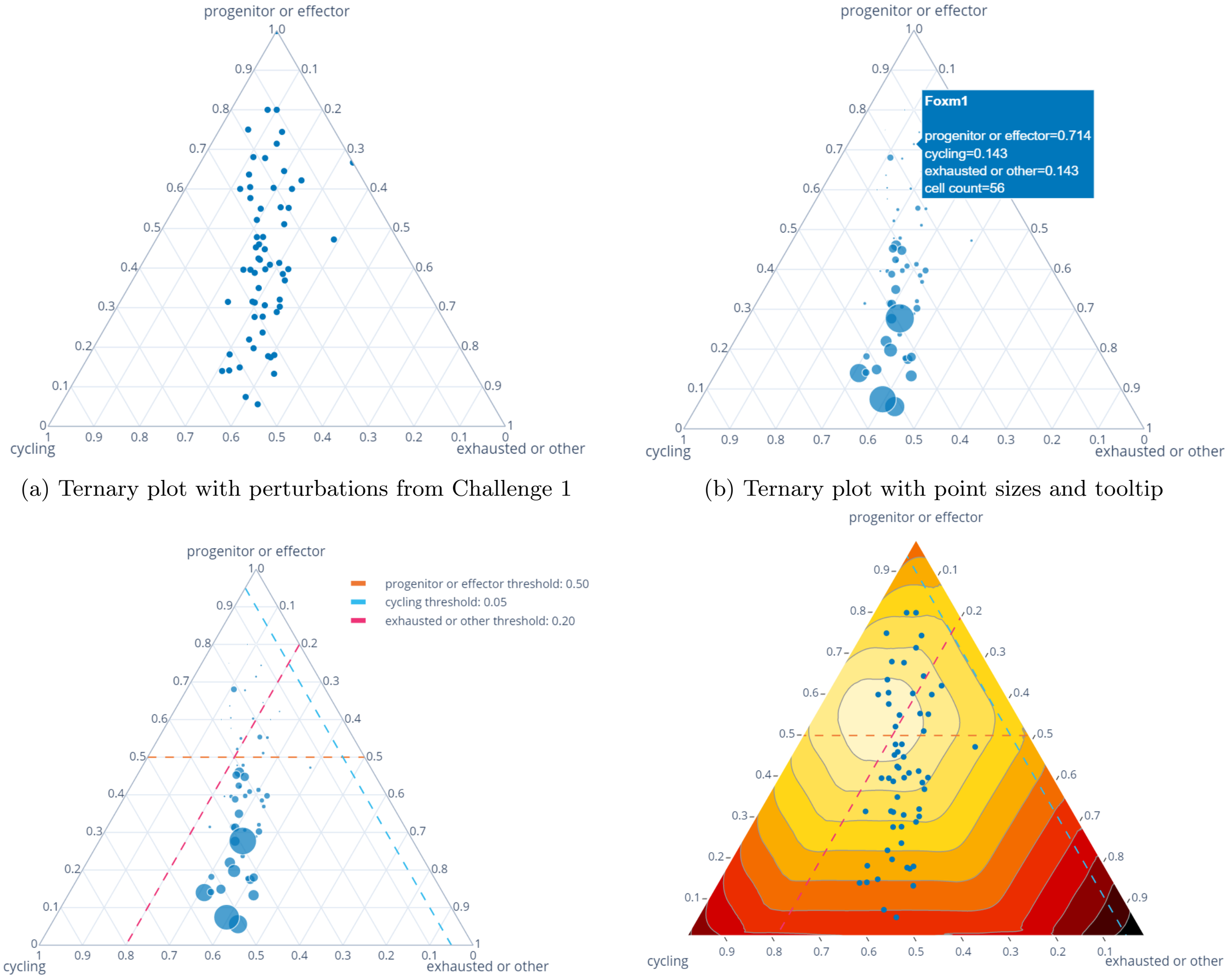
5 Visual assessment with interactive ternary plots

We believe that when working with large amounts of data, experts work best when equipped with interactive visual analysis tools. This is also the case with ranking cell perturbations, where the number of possibilities makes it difficult to make choices using solely tabular data representations. Therefore, in this section, we propose a simple interactive visualization that might be useful for analyzing cell states after knockouts (or predictions of knockouts) of thousands of genes.

As described in the challenge description, each analyzed cell is classified by experts into one of 5 cell states (progenitor, effector, terminal exhausted, cycling, other). These proportion vectors can also be thought of as probabilistic masses and will always add up to 1. That means that the values in cell proportion vectors are *sum-constrained*. Therefore, any four values in the proportion vector uniquely define the fifth value. This property allows visualizing any proportion vector using a *barycentric coordinate system*, which is tailored to sum-constrained data.

Unfortunately, 5-dimensional sum-constrained vectors need 4-dimensional visualizations. However, if some of the proportions were combined, one can visualize four proportion groups using a 3D tetrahedron visualization [7] or three proportion groups using a *ternary plot*. The latter visualization is presented in Figure 1. The code for generating the presented plots can be found in the Jupyter notebook submitted alongside this write-up.

[1] A more appropriate bound can be derived using combinatorial techniques for finite alphabets [5], but due to the brevity of this write-up we will use the Hoeffding bound, which gives a similar result and shows the same idea without as many equations.



(a) Ternary plot with perturbations from Challenge 1

(b) Ternary plot with point sizes and tooltip

(c) Ternary plot with proportion thresholds

(d) Ternary plot with a contour of an evaluation metric

Figure 1: Ternary plots for cell state proportion analyses.

In the ternary plots presented in Figure 1, each point represents a cell state proportion of a knockout from the Challenge 1 dataset. We combined the data to consist of vectors of 3 cell state groups: progenitor or effector (top vertex), cycling (lower left vertex), terminal exhausted or other (lower right vertex). One way of understanding this representation is to imagine a point in the ternary plot as the center of mass of the values in the proportion vector. If all cells are in the cycling state, then the entire mass of the predictions is at cycling, and the point coincides with vertex cycling. If all cells are progenitor or effector, the point will coincide with the progenitor or effector vertex. In other words, the higher the proportion of a given cell state, the closer the point is to the states vertex in the ternary plot.

There can be numerous variations to the ternary plot. Figure 1a shows a simple ternary plot with points representing cell state proportions of knockouts from Challenge 1. Figure 1b shows how cell counts can be visually encoded as point sizes and how the interactivity of the plot can provide additional information (e.g., the gene name[2]) about a given proportion. Figure 1c shows how certain thresholds, e.g. the requirement that a proportion has at least 0.05 cells in the cyclic state, can by visualized on the plots.[3] Finally, Figure 1d shows how the contours of evaluation metrics (in this case, κ_T) can be overlayed on the ternary plot to show the regions with the best proportions according to a given metric. We note that each plot can be interactively zoomed in to show just a subset of the data from a particular region.

Overall, we believe that interdisciplinary efforts benefit from interactive visualizations. In the case of genomics, AI predictions should usually be additionally verified by experts from a biological perspective. The presented visualization would facilitate such verification and hopefully bridge the gap between automated predictions and biology-based decision-making.

[2] A more elaborate implementation could provide a link to Uniprot, OpenTargets, or any other useful database.

[3] A slight extension of this functionality would be to color regions of interest defined by the thresholds.

References

[1] Cédric Villani. *Optimal Transport: Old and New*. Springer, 2008.

[2] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[3] Indrė Žliobaitė, Albert Bifet, Jesse Read, Bernhard Pfahringer, and Geoff Holmes. Evaluation methods and decision theory for classification of streaming data with temporal dependence. *Machine Learning*, 98:455–482, 2015.

[4] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

[5] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*, chapter LDP for Finite Dimensional Spaces, pages 11–70. Springer Berlin Heidelberg, 2010.

[6] Pedro Domingos and Geoff Hulten. Mining high-speed data streams. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 71–80, 2000.

[7] Dariusz Brzezinski, Jerzy Stefanowski, Robert Susmaga, and Izabela Szczech. Tetrahedron: Barycentric measure visualizer. In *Machine Learning and Knowledge Discovery in Databases*, pages 419–422, 2017.