GENVISAGE: Rapid Identification of Visually Explainable Features for Genomic Data Analysis

xxx
University of Illinois (UIUC)
xxx@illinois.edu

1. Introduction

2. Problem Definition

In Section 1, we have talked about different biology applications that can be abstracted as a separability problem. In this section, we will formally formulate the problem and discuss the challenges in it.

We first introduce some essential notations. Let $\mathcal{M}$ be a feature-sample matrix of size $m \times N$, where each row is a feature and each column is a sample. Correspondingly, denote the m features as $\mathcal{F} = \{f_1, f_2, \cdots, f_m\}$ and N samples as $\mathcal{S} = \{s_1, s_2, \cdots, s_N\}$. Each entry $\mathcal{M}_{i,j}$ in $\mathcal{M}$ is of numeric type, referring to the value of sample s_j on feature f_i. In addition, we are given two non-overlapping sets of samples, one with positive label and the other with negative label, denoted as $\mathcal{S}_+$ and $\mathcal{S}_-$ respectively. Both positive and negative samples are a subset of all samples, i.e., $\mathcal{S}_+ \subset \mathcal{S}$ and $\mathcal{S}_- \subset \mathcal{S}$. Let $\bar{\mathcal{S}}$ be the sample set with both positive and negative labels and n be the total number of samples, i.e., $\mathcal{S}_+ \cup \mathcal{S}_- = \bar{\mathcal{S}}$, $|\bar{\mathcal{S}}| = n$ and $n \leq N$. Furthermore, let l_k be the label of sample $s_k \in \bar{\mathcal{S}}$, i.e., $l_k = 1$ if s_k is positive and $l_k = -1$ if s_k is negative.

As illustrated in Figure 1, given matrix $\mathcal{M}$ and two sample sets $\mathcal{S}_+$ and $\mathcal{S}_-$, the goal is to find discriminative features to separate $\mathcal{S}_+$ from $\mathcal{S}_-$, and output a visualization to explain the separability. In GENVISAGE, we focus on finding TOP-K feature pairs instead of TOP-K single features. This is because *(a)* a single feature can be considered as a special case of a feature pair by taking the same feature in a feature pair; *(b)* a combined feature pair is likely to provide new insight compared to single features; *(c)* feature pair can be easily visualized as a 2-D space. As we will illustrate in Section 5, two features that perform poorly on their own may have very good separability when combined together. Furthermore, since we are targeting as a data exploration tool before more time-consuming machine learning methods, our design principle is to prioritize running time over accuracy. Next, let us formally define the *separability* problem.

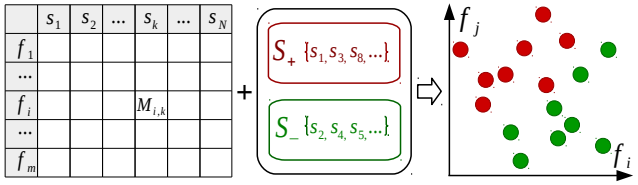


Figure 1: Different methods to Calculate Separability Score $\theta_{i,j}$

PROBLEM 1 (Separability). *Given a feature-sample matrix $\mathcal{M}$ and two labeled sample set $(\mathcal{S}_+, \mathcal{S}_-)$, **fast identify** TOP-K feature pairs (f_i, f_j) separating $\mathcal{S}_+$ from $\mathcal{S}_-$ based on a given **separability metric**, and output a 2-D **visualization**.*

There are three key aspects in Problem 1. First, the output of GENVISAGE is not just the TOP-K feature pairs, but also the corresponding visualizations. The output visualizations can help the users better interpret the result, i.e., how $\mathcal{S}_+$ is separated from $\mathcal{S}_-$. Second, given a feature pair (f_i, f_j), how to measure its quality in separating $\mathcal{S}_+$ from $\mathcal{S}_-$ poses an interesting challenge. Last but not least, how to fast identify TOP-K feature pairs based on the separability metric is also a big concern since small latency is critical for data exploration tool. In the following, we ellaborate more on these three points.

Visualization Output. Given a feature pair (f_i, f_j), it is natural to visualize the sample sets $(\mathcal{S}_+, \mathcal{S}_-)$ in a two dimensional space by coloring $\mathcal{S}_+$ as red and $\mathcal{S}_-$ as green, where the x-axis and y-axis represent feature f_i and f_j respectively. Thus, a visualization and a feature pair has an one-to-one relationship: each visualization corresponds to one feature pair and vice versa. By looking at the visualization, the users can easily have a general sense of how $\mathcal{S}_+$ and $\mathcal{S}_-$ are separated for a given feature pair, and have a better interpretation towards the results compared to the pure feature pair names.

Separability Metric. As far as we are concerned, existing separability measurements focus on single feature instead of feature pair. For instance, in order to characterize differentially expressed genes (DEG), biologists typically do association test (e.g, hypergeometric test) to find the best single feature that separate two gene sets. However, we argue that feature pair can provide new insights that is not revealed by top single features, as we will demonstrate in Section 5. Hence, developing a meaningful separability metric for feature pairs is the very first step towards our separability problem and is of great importance. Furthermore, since GENVISAGE has visualization as the output instead of a pure separability score, our proposed separability metric should also encodes the visual separability to some extent.

Fast Identification. Since GENVISAGE serves as a data exploration tool before looking into more sophisticated machine learning algorithms, we in particular care more about the running time than the accuracy. For one thing, instead of complicated machine learning methods, we prefer light-weight separability metrics. In some sense, we try to solve Problem 1 in a "quick and dirty" way. For another, various optimization mechanisms are essential to further cut down the running time.

In the following, we will first describe our selected separability metric in Section 3, and then discuss different optimization techniques in Section 4.

3. Separability Metric

As discussed in Section 2, given a feature pair (f_i, f_j) we can visualize $\mathcal{S}_+$ and $\mathcal{S}_-$ in a 2-D space. In the following, we propose a separability metric based on linear separability [2] for a 2-D visualization.

Let us first review the concept of *linear separability* introduced in Euclidean geometry [2]. A pair of point sets in two dimensions, i.e., $\mathcal{S}_+$ and $\mathcal{S}_-$, are *linearly separable* if there exits at least one straight line in the plane such that all points from $\mathcal{S}_+$ are on one side of the line, while all points from $\mathcal{S}_-$ are on the other side of the line. Mathmatically, we can represent a line ℓ using Equation 1, where x and y represent a sample's value on feature f_i and f_j respectively, and w_0, w_i and w_j are coefficients. Given a feature pair (f_i, f_j) and a line ℓ, let $\eta_{i,j}^{\ell,k}$ be the estimated label of a sample s_k. The estimated label $\eta_{i,j}^{\ell,k}$ is calculated according to Equation 2: if s_k lies on the upper side of ℓ, then $\eta_{i,j}^{\ell,k} = 1$; otherwise, $\eta_{i,j}^{\ell,k} = -1$. If there exists a line ℓ such that for any samples $s_k \in \bar{\mathcal{S}}$, the estimated label $\eta_{i,j}^{\ell,k}$ is consistent with the real label l_k as shown in Equation 3, then we say $\mathcal{S}_+$ and $\mathcal{S}_-$ are linear separable.

$$\ell: \quad w_i \cdot x + w_j \cdot y + w_0 = 0 \tag{1}$$

$$\eta_{i,j}^{\ell,k} = \text{sign}(w_i \cdot \mathcal{M}_{i,k} + w_j \cdot \mathcal{M}_{j,k} + w_0) \tag{2}$$

$$\eta_{i,j}^{\ell,k} = \begin{cases} 1 & \textit{if} \quad s_k \in \mathcal{S}_+, \textit{ i.e., } l_k = 1 \\ -1 & \textit{if} \quad s_k \in \mathcal{S}_-, \textit{ i.e., } l_k = -1 \end{cases} \tag{3}$$

Next, let us introduce our proposed separability metric. Our proposed separability metric is based on linear separability and is defined as *how well a 2-D visualization can be linearly separated*. This is because GENVISAGE stresses on visual explainability and linear separability in a 2-D visualization can be easily recognized and interpreted by the users. We then formally define the separability metric. First, given a feature pair (f_i, f_j) and a line ℓ in the 2-D plane, a sample s_k is said to be correctly separated if Equation 3 holds, i.e., $\eta_{i,j}^{\ell,k} \cdot l_k = 1$. Correspondingly, let $\tau_{i,j}^{\ell,k}$ be the variable indicating whether sample s_k is correctly separated, as depicted in Equation 4. Then, given a feature pair (f_i, f_j) and a line ℓ, the separability score is defined as the number of correctly separated samples, denoted as $\theta_{i,j}^{\ell}$ as shown in Equation 5. Figure 2a shows some possible $\theta_{i,j}^{\ell}$ with different separating lines. Finally, the separability score for a feature pair (f_i, f_j) is defined as the largest $\theta_{i,j}^{\ell}$ among all possible lines ℓ, denoted as $\theta_{i,j}$ as shown in Equation 6.

$$\tau_{i,j}^{\ell,k} = \begin{cases} 1 & \textit{if} \quad \eta_{i,j}^{\ell,k} \cdot l_k = 1 \\ 0 & \textit{otherwise} \end{cases} \tag{4}$$

$$\theta_{i,j}^{\ell} = \sum_k \tau_{i,j}^{\ell,k} \tag{5}$$

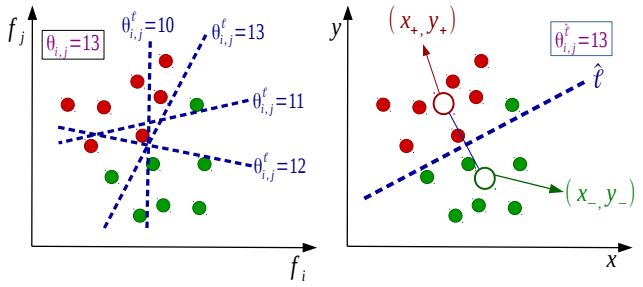
$$\theta_{i,j} = \max_{\ell} \{\theta_{i,j}^{\ell}\} \tag{6}$$

Brute Force. As illustrated in Figure 2a, a brute force way to calculate $\theta_{i,j}$ is to first enumerate all possible separating lines ℓ and calculate each $\theta_{i,j}^{\ell}$. This is infeasible as there are infinite number of possible lines. However, we can easily trim down the search space to $O(n^2)$ lines by linking every two points in the 2-D plane. This is because the results of all other possible lines can be covered by these $O(n^2)$ lines. Nevertheless, it is still very time-consuming to consider $O(n^2)$ lines for each feature pair (f_i, f_j). We further propose to use a *representative line* $\hat{\ell}$ to approximate $\theta_{i,j}$ as shown in Equation 7, instead of considering all $O(n^2)$ possible lines and pick the best in Equation 6. Now, the search space is reduced to $O(1)$ from $O(n^2)$. The representative line is picked based on Rocchio's algorithm [1]. As we will show later in Section 5, $\theta_{i,j}^{\hat{\ell}}$ is comparable to $\theta_{i,j}$ when using Rocchio-based representative line $\hat{\ell}$. Let us describe in detail about the Rocchio-based representative line.

$$\theta_{i,j} \approx \theta_{i,j}^{\hat{\ell}} \tag{7}$$

Rocchio-based. The basic idea of Rocchio's algorithm is to estimate each sample's label the same as its nearest centroid. More specifically, let us denote the centroid of positive samples $\mathcal{S}_+$ and negative samples $\mathcal{S}_-$ as $\mu_+ = (x_+, y_+)$ and $\mu_- = (x_-, y_-)$ respectively ,as shown in Figure 2b. Link μ_+ with μ_-, then the perpendicular bisector is defined as the representative separating line $\hat{\ell}$ as illustrated in Figure 2b. Mathmatically, $\hat{\ell}$ can be represented as Equation 8 with instantiated w_i, w_j and w_0 in Equation 1. The corresponding $\theta_{i,j}^{\hat{\ell}}$ equals 13 with one negative sample (blue point in Figure 2b) mis-estimated as positive.

$$\hat{\ell}: \quad (x_+ - x_-) \cdot x + (y_+ - y_-) \cdot y - \left(\frac{x_+^2 - x_-^2}{2} + \frac{y_+^2 - y_-^2}{2}\right) = 0 \tag{8}$$



(a) Brute Force (b) Rocchio-Based

Figure 2: Different methods to Calculate Separability Score $\theta_{i,j}$

Brute-force v.s. Rocchio-based. Compared to brute force, Rocchio-based method is more light-weight in terms of running time, but at the cost of accuracy in calculating $\theta_{i,j}$. Intuitively, Rocchio-based representative line is a reasonable approxy to the best separating line since Rocchio-based method assigns each sample to its nearest centroid and each centroid is calculated as the representative of each group. We will further experimentally demonstrate that $\theta_{i,j} - \theta_{i,j}^{\hat{\ell}}$ is small in Section 5.

4. Proposed Optimizations

5. Experiment

References

[1] J. J. Rocchio. Relevance feedback in information retrieval. 1971.

[2] M. I. Shamos. Geometric complexity. In *Proceedings of seventh annual ACM symposium on Theory of computing*, pages 224–233. ACM, 1975.