Mathematical formula for the metric Relative expression across cell type clusters

Anonymous Authors[1]

Abstract

This document provides the suggestion of Asli and Katelyn for the Relative expression across cell type clusters metric. The process for the similarity metric: Relative expression levels/probe efficiency is the same by converting the pairwise differences per gene for each cell to pairwise differences per cell for each gene.

0.1. Relative expression across cell type clusters

The similarity metric, Relative expression across cell type clusters, measures how similar the relative expression between cell types are within each gene of single cell dataset and the spatial dataset.
First, it computes a similarity value for each gene across clusters which is measured by the difference between mean-normalized gene expression of genes across clusters in both modalities.
Later, one can use this similarity value for different functionalities in metric. Functionalities are to compute the mean, median or the percentile 95 similarity of the difference between mean normalized expression of genes across clusters.

0.1.1. Calculate the difference between mean normalized expression of genes across clusters in both modalities

Let k be the number of genes and l the number of cells within cell type clusters. Let $U = \{U_1, ..., U_n\}$, $V = \{V_1, ..., V_n\}$ be the set of sets of cells in each cell type cluster for single-cell data and the spatial data. We have $|U| = |V| = n$ number of genes.
Denote the shared genes as I where $I = U \cap V$ with $|I| = n$ and subset the single-cell data U to spatial data V with $\hat{U} = U \cap I$ to include only the unique cell types. We have $|\hat{U}| = m$

Define the log-normalized gene expression level of gene I_i and denote as $\exp_{*,\mathrm{k,i}}$ where $i = \{1, \ldots, n\}$ denotes gene, $k = \{1, \ldots, m\}$ the related cell and $* := \{sc, sp\}$ representing the single-cell or the spatial data.

Define the mean gene expression levels for each cell type as $Mean_{*,i,j}$ for both modalities by dividing the log-normalized gene expression level of gene i in unique cell j:
$Mean_{*,i,j} = \frac{\left(\sum_{j=0}^{m} \exp_{*,\mathrm{i,j}}\right)}{m}$

(**from here the code and the suggestion differs**)

Define mean-normalized gene expression levels for each cell type by dividing each mean gene expression value by the corresponding mean value of the whole gene i for every cell j:
$Mean_{norm,*,i,j} = \frac{Mean_{*,i,j}}{\sum_{s=1}^{m} \frac{Mean_{*,i,s}}{m}}$

Then define the mean absolute difference value between the mean-normalized gene expression levels and each cell type which represents the similarity values S_i with $i = \{1, \ldots, n\}$ for each gene for the next subsection:

$Mean_{abs,i} = \frac{|Mean_{norm,sc,i,j} - Mean_{norm,sp,i,j}|}{|Mean_{norm,sc,i,j} - Mean_{norm,sp,i,j}|}$

Denote mean absolute difference as S_i for gene i from now on.

0.1.2. Mean similarity, Median similarity and Percentile 95

Compute mean similarity of the difference between mean normalized expression of genes across clusters in both modalities by taking the mean of the similarity values:
Mean similarity $= \frac{\sum_{i=1}^{n} S_i}{n}$
Similarly *Median* representing the middle value of the similarity values and *percentile 95 rank* for statistical computations.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

(**not in the code - but a suggestion**)
Calculate the pairwise difference between cell types for each column using the following idea:

For each unique pair of columns (col_i, col_j) from $Mean_*$ and such that $i < j$ compute the difference in between and store the values in P_* which represents the pairwise differences per gene for single cell data and analog for the other modality:

$P_* = (p_{k,j})$ with $(p_{k,j})_{1 \leq k \leq \frac{m!}{(m-2)!*2!}, 1 \leq j \leq n}$

where m= Number of cells and n= Number of genes in M_*.
For $Mean_* = (m_{i,l})_{1 \leq i \leq m, 1 \leq l \leq n}$ we have

$p_{k,j} = \sum_{k=1}^{\frac{m*(m-1)}{2}-1} \sum_{j=i+1}^{n} [Mean_{j,col_i} - Mean_{k,col_l}]$

and analog for P_{sp}.

To define the mean normalized relative expression matrix for each gene, we define:

$$\mathfrak{Mean}^*_{i,j} = \left(\frac{p_{i,j}}{\sum_{s=1}^{\frac{m*(m-1)}{2}} p_{s,j}} \right)_{1 \leq i \leq \frac{m*(m-1)}{2}, 1 \leq j \leq n}$$

Analog for $\mathfrak{M}^{sp}_{i,j}$.

As last step we define $S_i = \frac{|\mathfrak{Mean}^{sc} - \mathfrak{Mean}^{sp}|}{|\mathfrak{Mean}^{sc} - \mathfrak{Mean}^{sp}|}$