MADVAR: An Algorithm that Improves the Relevance of Computational Biology Output While, Reducing Compute Time and Space Requirements

Gilad Silberberg, PhD and Michael Ritchie, PhD, MBA

Champions Oncology, Inc and Corellia AI, Inc, Rockville, MD, USA

ABSTRACT

High-throughput methods implemented in biology research produce a continuously growing array of data input used to produce data output with an increasing abundance of features. While growth in the volume and diversity of data input can be highly valuable for studying biological systems, it presents the challenge of managing enormous quantities of features, many of which are not relevant to the specific research question being asked. This excess data input burdens storage and computation of downstream clustering and machine learning (ML) tasks. A common approach used to manage this data input relies on filters applied to the features by their variance across the sample set, while applying random cutoffs.

Our proprietary algorithm, MADVAR, enables the prioritization of variable features from high-throughput continuous data, by automatically finding an optimal cutoff for the distribution of the data. Based on the right-skew nature of biological data variance distribution, MADVAR finds and excludes the "0 variance peak" using the median of the distributions and the median absolute deviation (MAD). MADVAR enables a faster analysis with a reduced memory requirement, and dramatically improves clustering results with minimal loss of relevant features.

MATERIALS & METHODS

MADVAR cutoff is calculated as median + B*MAD, where the coefficient B = 2 in this study, but can be adjusted by the user. Alternatively, the function allows to use mode*B as a cutoff. Euclidean distance method was used throughout the study.

The Ward.D method was used for hierarchical clustering. Random forest was run with ntree > = 250.

RESULTS

To demonstrate the assumption of right-skewed variance distribution in various biological dataset, we selected multiple data sets of diverse types, including RNAseq read counts (fig. 1A-C) and RNAseq TPM (fig. 1D) and proteomics (fig. 1E-G), and observed their variance distribution densities. Indeed, all data types displayed similar distribution shapes, with a high peak on the left end corresponding to near-zero variance. Consistently, the mode of the variance corresponded to the summit of the peak, while the median was immediately after, arguably at the center of the peak (blue and gray dashed lines, respectively). The MADVAR cutoff was calculated as median + 2*MAD (red line). It is shown that MADVAR can consistently and accurately identify the "0 variance peak".

To assess the performance of MADVAR, two additional, commonly applied cutoffs were used as a benchmark (Table 1). The differentially filtered datasets were then clustered by several methods, both unsupervised and supervised. Subsequently, the quality and performance of the clustering was measured and compared.
To assess and compare the quality of unsupervised clustering, we calculated the connectivity, Dunn index and Homogeneity index (BHI) in the different datasets (fig.2-4, respectively), where the number of clusters (k) was applied according to the number of levels in each grouping term, shown in table 2. Clustering of MADVAR-filtered datasets provided the most connected (fig. 2) and phenotypically homogeneous (fig. 4) clusters, more frequently than the other filtering approaches. While MADVAR was not the most frequent best performer according to the Dunn index (which measures the inter- to intra cluster distance ratio), in specific cases it has dramatically better scores than the other approaches (fig. 3 Breast and NSCLC RNA).

To compare classification performance of supervised learning, we chose random forest as a representative ML method. It was run in multiple seed iterations (n = 48), from which the out-of-bag (OOB) error rates were extracted and used as a performance metric. Here too, MADVAR showed the best performance, having the lowest mean or median error rate in the majority of datasets (fig. 5).

RESULTS

Dataset	Type	# Samples	Above 20%	Above 80%	MADVAR	Grouping
TCGA	RNAseq normalized counts	1312	38502	9626	11000	Clinical status, Tissue, Primary disease
CO SARCOMA	RNAseq TPM	143	10840	2710	3206	Tumor subtype
CO SARCOMA	Proteomics (relative abundance)	135	2737	685	579	Tumor subtype
CO BREAST	RNAseq TPM	131	10379	2595	2974	Tumor status
CO BREAST	Proteomics (relative abundance)	122	4017	1005	720	Tumor status
CO NSCLC	RNAseq TPM	200	10595	2649	3134	Histology
CO NSCLC	Proteomics (relative abundance)	148	3581	896	640	Histology

Table 1. Summary of the datasets used in the study. “TCGA” included indications for which normal samples were available. CO: Champions Oncology datasets. TPM: Tags per million. Columns “Above 20%”, “Above 80%”, “MADVAR” show the number of features left after applying the corresponding filter.

Grouping	Cohort	Groups	N
sample_type	TCGA	Normal, Tumor	2
primary_disease	TCGA	bladder_urothelial_carcinoma, breast_invasive_carcinoma, colon_adenocarcinoma esophageal_carcinoma, head_neck_squamous_cell_carcinoma, kidney_chromophobe, kidney_clear_cell_carcinoma, kidney_papillary_cell_carcinoma, liver_hepatocellular_carcinoma, lung_adenocarcinoma, lung_squamous_cell_carcinoma, prostate_adenocarcinoma, rectum_adenocarcinoma, stomach_adenocarcinoma, thyroid_carcinoma, uterine_corpus_endometrioid_carcinoma	16
Tissue	TCGA	Breast, colorectal, head_neck, kidney, liver, lung, prostate, stomach, thyroid, other	10
Tumor subtype	Champions Oncology	Sarcoma: EWING SARCOMA, GIST, LEIOMYOSARCOMA LIPOSARCOMA, OSTEOSARCOMA, Other	6
Tumor subtype	Champions Oncology	Breast: METASTATIC, PRIMARY	2
Histology	Champions Oncology	NSCLC: ADENOCARCINOMA, SQUAMOUS CARCINOMA, Other	3

Table 2. Summary of phenotypic grouping used in the study. The number of levels (N) was used as k in unsupervised clustering assessment. GIST: GASTROINTESTINAL STROMAL TUMOR

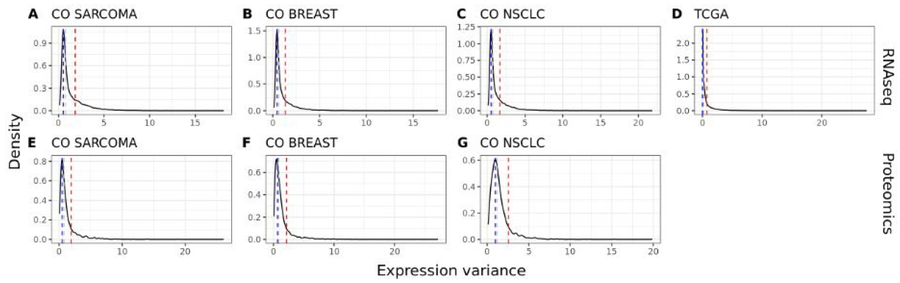

Figure 1. Distribution densities of various data types (see table 1). Blue: mode (Asselin estimate, bw = 0.9). Gray: median. Red: final cutoff (median + MAD * mads), where mads = 2.

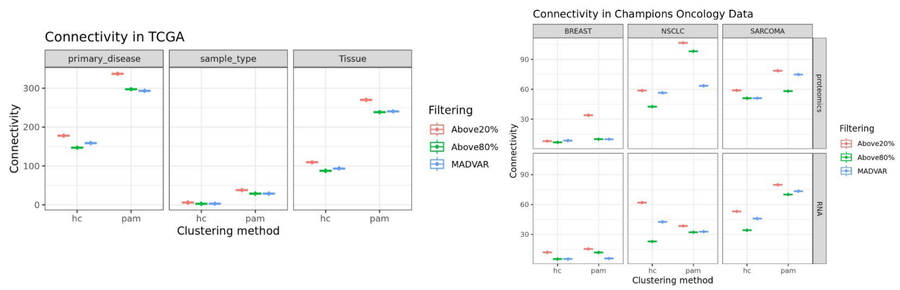

Figure 2. The connectivity indicates the degree of connectedness of the clusters, as determined by the k-nearest neighbors. The connectivity has a value between 0 and infinity and should be *minimized*. hc: hierarchical clustering. Pam: partition around medoids

RESULTS

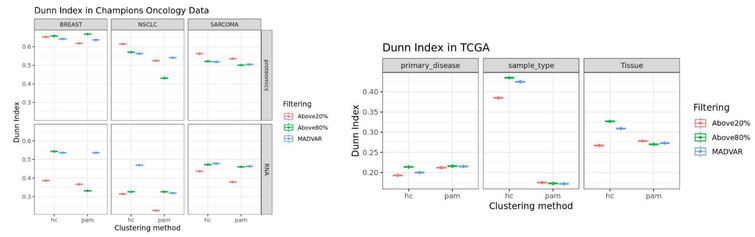

Figure 3. The Dunn Index is the ratio of the smallest distance between observations not in the same cluster to the largest intra-cluster distance. The Dunn Index has a value between zero and infinity and should be *maximized*. hc: hierarchical clustering. Pam: partition around medoids

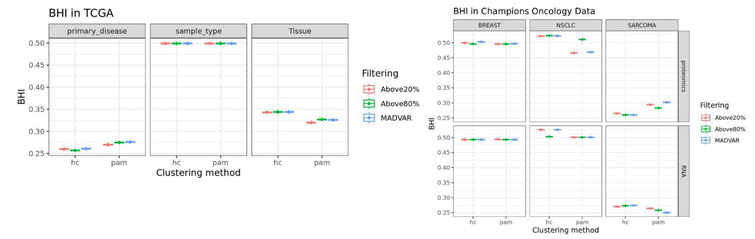

Figure 4. Homogeneity index (BHI). Measures the average proportion of sample pairs that are clustered together which have matching annotation classes. The BHI is in the range [0,1], with larger values corresponding to more homogeneous clusters. hc: hierarchical clustering. Pam: partition around medoids

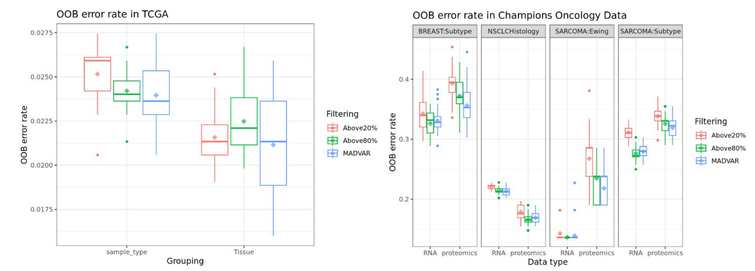

Figure 5: Out-of-bag (OOB) error rates measured in multiple runs of random forest (n=48). The out-of-bag error is a performance metric that estimates the performance of the Random Forest model using samples not included in the bootstrap sample for training. The diamond inside the box indicates the mean.

SUMMARY

MADVAR enables a faster analysis with a reduced memory requirement, while improving clustering results with minimal loss of relevant features.

Conceptually, it optimizes the balance between feature selection, an essential step in ML approaches, and the integrity of data elements required to explain a complete biological system.

SCAN ME