An Empirical Study of Cluster Evaluation Metrics using Flow Cytometry Data

Nima Aghaeepour *
Terry Fox Laboratory
BC Cancer Agency
675 West 10th Avenue
Vancouver, BC, V5Z1L3, Canada
naghaeep@bccrc.ca

Alireza Hadj Khodabakhshi *
Terry Fox Laboratory
BC Cancer Agency
675 West 10th Avenue
Vancouver, BC, V5Z1L3, Canada
akhodaba@bccrc.ca

Ryan R. Brinkman
Terry Fox Laboratory
BC Cancer Agency
675 West 10th Avenue
Vancouver, BC, V5Z1L3, Canada
rbrinkman@bccrc.ca

Abstract

A wide range of abstract characteristics of partitions have been proposed for cluster evaluation. We empirically evaluated the performance of these metrics for flow cytometry data and found that the set-matching metrics perform closest to human.

1 Introduction

Clustering is an increasingly popular module in data processing applications. Many clustering algorithms have been developed and many more are anticipated to emerge in the future. Thus, methods for assessing the performance of a clustering algorithms are in great demand. Such methods assess the performance of a clustering algorithm by computing a *quality score* of the solution against a *ground truth* partition, usually designed by a human expert. A wide range of these *criterion* have been proposed [9]. Evaluating clustering algorithms heavily relies on the chosen quality score however, it is not practical to study the performance of these metrics in a domain-independent way [3]. In this paper we aim to empirically evaluate the available metrics to find the best metric for comparing clustering solutions against ground truth partitions for flow cytometry (FCM) applications. This work was motivated in part by the challenges we faced in choosing the best clustering comparison metric for the FlowCap project. FlowCap is an international open project designed to provide an objective way to compare and evaluate FCM data clustering methods, and also to establish guidance about appropriate use and application of these methods (for more information visit http://flowcap.flowsite.org/).

2 Background

We studied three representatives of three categories of clustering comparison metrics: (1) pair counting ([7]), (2) set matching ([6, 10]) and (3) entropy based ([6, 9]). We will refer to the gold standard partitions as CLASSES (C) and to hypothesized partitions as CLUSTERS (K).

All Pair Counting Measures: The all-pair counting measures evaluate the similarity between two clustering results by examining how likely they are to group a pair of data points together, or, sepa-

*These authors contributed equally to this work.

rate them in different clusters. *Rand Index* [7] is defined as the percentage of the pair of data points that are either clustered together in both C and K or clustered differently in both C and K.

Set Matching Measures: Set matching measures seek for a match between clusters and classes and compute the measure based on the overlap between the matching clusters/classes. While *Misclassification Rate (MCR)* [2] exhaustively searches for a best match between cluster and classes, *F-measures* [8] greedily search for a best match for each cluster and adds up these contributions.

Entropy-Based Measures: Entropy based measures use tools from information theory, used to compare two clustering solutions. V-measure [9] measures how successfully the criteria of *homogeneity* and *completeness* have been satisfied. A clustering result is homogeneous (complete) if all of its CLUSTERS (CLASSES) contain only data points which are members of a single CLASS (CLUSTER). V-measure is defined as weighted harmonic mean of homogeneity and completeness. Variation of Information (VI) [6] measures the amount of information lost and gained in changing from clustering C to K. This metric is defined as the sum of conditional entropy of the clustering K given the class C and the conditional entropy of C given K.

3 Data Sets

We used two independent FCM datasets: Diffuse Large B-cell Lymphoma **(DLBCL)** [4] and Graft versus Host Disease **(GvHD)** datasets [1]. 30 DLBCL and 12 GvHD samples were manually analyzed by human experts and their cell populations (i.e. clusters) are identified. The DLBCL dataset is known to have two cell populations so it was clustered using a t-distribution mixture model [5] with the number of clusters set to two. The number of clusters in the GvHD dataset is unknown so flowmerge [2] was used for choosing the number of clusters.

4 Methodology

To provide a gold standards for the metrics, for each sample, we visually compared the automatic and manual clusters and labeled each sample as **success** or **failure** when the two clustering solutions are similar or dissimilar respectively. Figure 1 shows an example of success and failure samples.

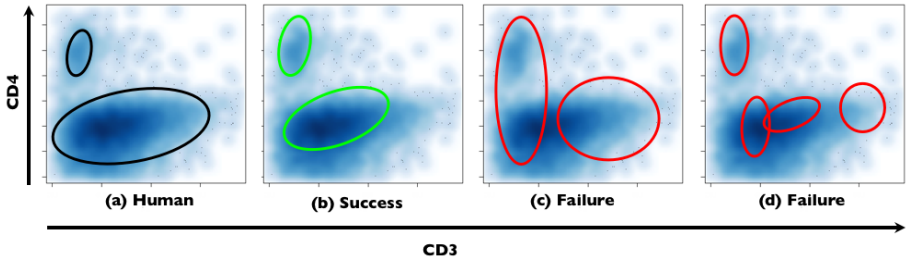

Figure 1: Four different clustering solutions. (a) is identified by a human expert; (b), (c) and, (d) are predicted using automated clustering algorithms. (b) is labeled a as success (*i.e.*, similar to human); (c) is labeled as a failure because the shape of the clusters are different from the humans output; (d) is failure because of the number of predicted clusters is high.

We analyzed the metrics to find the one that performs the best in classifying success and failure samples. We normalized the output of the metrics to fall in $[0, 1]$ interval. Hence, a "good" metric for cluster evaluation is a metric that returns a value close to 0 for failure samples and a value close to 1 for success samples. In other words, we expect the distribution of metric values to be skewed to the left (close to 0) and right (close to 1) for failure and success samples, respectively.

5 Results

Figure 2 shows the estimated distribution of the metrics. This figure shows the metrics that are from the same family are correlating with each other. For a more detailed comparison, in Figure3

we picked a representative from each family and compare their Cumulative Distribution Functions (CDF). For failure samples (Figure 3(1)) F-measure has a better performance and for success samples (Figure 3(b)) V-measure has a small advantage. In addition, in Figure 4 we provided a quantitative measure by showing the average distance of each metric to the optimum measure (i.e., 0 for failure and 1 for success samples). On average, F-measure has the minimum overall error while Homogeneity has the best performance in discriminating failure cases.

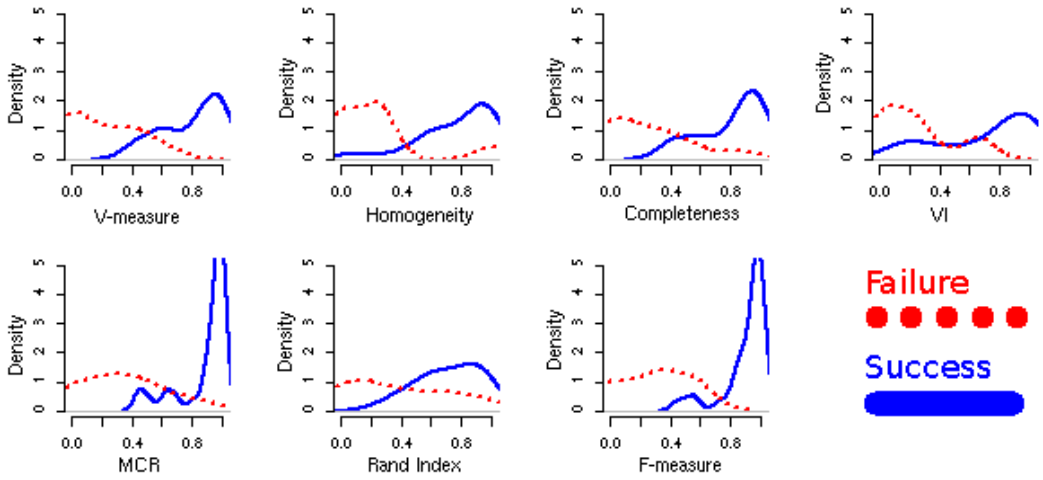

Figure 2: Kernel Density Estimation (KDE) of 7 metrics.

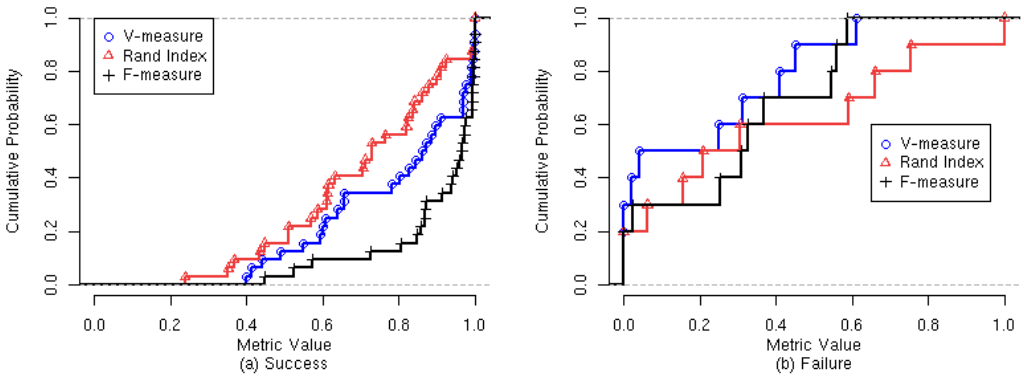

Figure 3: CDF of (a) failure and (b) success samples for 3 metrics.

6 Discussion and Conclusion

Our study shows that the metrics are not strongly correlating for evaluating FCM data clustering algorithms. The set matching measures (i.e. MCR and F-Measure) perform closer to human. A possible reason for this result is that the human expert tries to match the clusters when evaluating the algorithm. However, this does not necessarily mean that the set matching metrics are better choices for cluster evaluation. It only means that the set matching measures perform similar to FCM experts. Finally, V-Measure shows an acceptable performance specially in discriminating failure

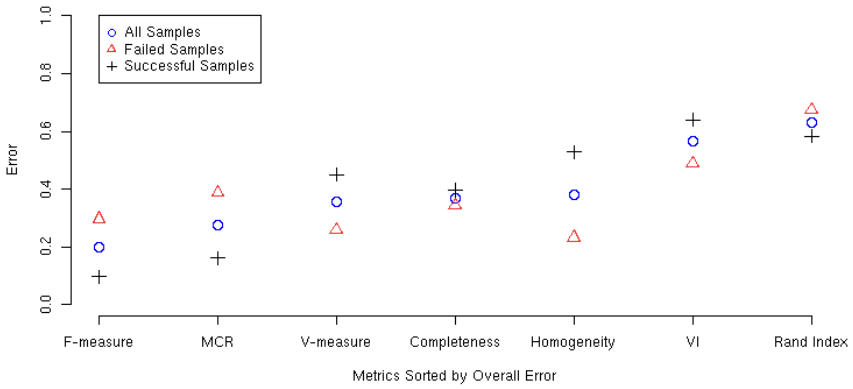

Figure 4: Single value representatives for the distance of each metric to the optimum measure (i.e., 0 for failure and 1 for success samples).

samples. Having a distribution with only two major modes is an important advantage specially when the user needs to use thresholding strategies for automatic clustering algorithm evaluation.

7 Acknowledgements

This project was supported by a Michael Smith Foundation for Health Research Scholar Award to RRB, a MSFHR/CIHR scholarship to NA, and by NIH grant 1R01EB008400.

References

[1] R. R. Brinkman, M. Gasparetto, S. J. Lee, A. J. Ribickas, J. Perkins, W. Janssen, R. Smiley, and C. Smith. High-content flow cytometry and temporal data analysis for defining a cellular signature of graft-versus-host disease. *Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation*, 13(6):691–700, Jun 2007.

[2] Finak G, Bashashati A, Brinkmann R, and Gottardo R. Merging mixture model components for improved cell population identification in high throughput flow cytometry data. *Advances in Bioinformatics*, page to appear, 2009.

[3] Isabelle Guyon, Ulrike von Luxburg, and Robert C. Williamson. Clustering: Science or art? *NIPS 2009 Workshop on Clustering Theory, Vancouver, Canada*, 2009.

[4] F. Hahne, A. Hadj Khodabakhshi, A. Bashashati, CJ. Wong, RD. Gascoyne, AP. Weng, S. Seifert-Margolis, A. Bourcier, K.; Asare, T. Lumley, R. Gentleman, and RR. Brinkman. Per-channel basis normalization methods for flow cytometry data. *to appear in Cytometry A*.

[5] Kenneth Lo, Florian Hahne, Ryan Brinkman, and Rafael Gottardo. flowclust: a bioconductor package for automated gating of flow cytometry data. *BMC Bioinformatics*, 10(1):145, 2009.

[6] M. Meila. Comparing clusterings-an information based distance. *Journal of Multivariate Analysis*, 98:873–895, 2007.

[7] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66:846–850, 1971.

[8] C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 1979.

[9] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. *In Proc. of the 2007 Joint Conference on Empirical Methods in NLP and Computational Natural Language Learning*, pages 410–420, 2007.

[10] Ding Zhou, Jia Li, and Hongyuan. A new mallows distance based metric for comparing clusterings. *In Proc.22nd internal conference on Machine Learning*, pages 1–8, 2005.