Multi-objective Evolutionary Clustering of Single-Cell RNA Sequencing Data

Konghao Zhao
Wake Forest University
Winston Salem, North Carolina, USA
zhaok220@wfu.edu

Natalia Khuri
Wake Forest University
Winston Salem, North Carolina, USA
natalia.khuri@wfu.edu

ABSTRACT

Cells are the basic building blocks of human organisms. Single-cell RNA sequencing is a technology for studying the heterogeneity of cells of different organs, tissues, conditions, and treatments. Identification of cell types and states in sequenced data is an important and challenging task, requiring computational approaches that are accurate, robust, and scalable. Existing approaches use cluster analysis as the first step of cell-type prediction. Their performance remains limited because they optimize only one objective function. In this study, two evolutionary clustering approaches were designed, implemented, and systematically validated, namely a single-objective evolutionary algorithm and a multi-objective evolutionary algorithm. Performance of the algorithms was validated on synthetic and experimental datasets. Both evolutionary algorithms were accurate, consistent, stable, and on par with or better than baseline algorithms. Run-time analysis of multi-processed implementation on an High Performance Computing Cluster, demonstrated that both evolutionary algorithms efficiently handle large datasets.

CCS CONCEPTS

• **Computing methodologies** → **Genetic algorithms**; • **Applied computing** → **Bioinformatics**.

KEYWORDS

bioinformatics, clustering, evolutionary algorithms, multi-objective optimization

1 INTRODUCTION

Due to powerful sequencing technologies, it is now possible to measure gene expression of individual cells and catalog their types and states. These catalogs, known as cell atlases, may be used to diagnose, treat, and monitor human diseases and disorders [10].

Identification of cell types is based on prior knowledge about the marker genes, which are differentially expressed in specific cell types and conditions. To identify cells with similar markers, they must be clustered into groups first, such that the gene expressions of cells within one group are similar and the gene expressions of cells in different groups are dissimilar. Despite these two objectives for cell-type clustering, most current approaches optimize for one objective only. For example, the well-known partitioning algorithm, Kmeans, optimizes the total sum of intra-cluster distances [9], whereas another widely used algorithm, PhenoGraph, optimizes clusters' modularity [8].

Interestingly, the concept of a cluster can not be precisely defined, and the true cluster memberships are unknown. Therefore, even though a clustering algorithm may produce the most compact clusters, the results may still be unusable because the clusters are not well separated. Poorly separated clusters hinder the identification of marker genes that are differentially expressed in neighboring clusters.

On the other hand, a clustering problem can be defined as a multi-objective optimization (MOO) problem. In MOO, the objectives are usually conflicting, and therefore, no single solution exists that simultaneously optimizes all of the multiple objective functions. Several methods may be used to solve the MOO problems. For example, evolutionary algorithms (EAs) have been proven mathematically to find an accurate approximation to the solutions of MOO problems [2]. EAs were inspired by the process of biological evolution, which uses the concept of the survival of the fittest [4, 6]. The best solution is found by the iterative improvement of an initial population of solutions using genetic operators of selection, crossover, and mutation.

2 DATA AND METHODS

In this work, each clustering solution was encoded by the real-valued chromosomes of length $K \times N$, where K is the number of user-defined clusters and N is the dimension of the dataset. EA parameters were experimentally determined using well-established clustering benchmarks from KEEL repository [1]. Specifically, to generate offsprings from the parents' population, we used a one-point crossover, polynomial mutation, and a tournament selection with a 0.8 crossover rate, 0.07 mutation rate, 1.0 individual mutation rate. The tournament size was set to 20% of the population, determined in small experiments. Additionally, in multi-objective evolutionary algorithm (MOEA), we performed a non-dominated sorting of parents' and offsprings' pools and selected a new population from the sorted pool using their non-dominance status [3, 7].

The internal and external validity of the two algorithms were extensively studied with 6 synthetic and 48 real single-cell RNA-sequencing (scRNA-seq) datasets. Each dataset comprised between 50 and 10,000 cells and between 10,000 and 32,502 genes. On average, each dataset comprised more than 50% of zero-inflated expression values. To preprocess the data, we filtered, log-transformed, scaled and reduced the dimensionality of each dataset. The datasets' size, diversity, dimensionality, and complexity allowed us to test and compare our algorithms rigorously.

We compared single-objective evolutionary algorithm (SOEA) and MOEA to existing scRNA-seq clustering methods, such as Kmeans and PhenoGraph. The efficiency of EAs was examined on a High-Performance Computing (HPC) cluster with 94 computing nodes, 4,224 processor cores, 68,608 GPU cores, 18.67TB memory,

and 221TB storage [11]. Run-time analysis of multi-processed EAs consumed a total of 4 nodes and 128 CPUs with 64GB memory.

SOEA and MOEA were implemented in Python using DEAP package (version 1.3.1) [5]. Data collection and preprocessing were performed using scripts written in R programming language.

3 RESULTS

Experimental results on scRNA-seq data, demonstrated that SOEA and MOEA clusterings were accurate, consistent, stable, and efficient. We evaluated internal and external clustering validity of the algorithms and compared them to two other clustering methods. Internal validity evaluation was performed on 48 real scRNA-seq datasets without unknown cluster memberships. We evaluated external validity of the algorithms on 6 synthetic datasets with the known number of clusters.

Result 1: The internal and external validity of MOEA and SOEA were on par with or better than those of the baseline algorithms. Internal validity, measured using Silhouette Coefficient (Sil), was 0.485 for MOEA and 0.442 for SOEA, averaged over 48 experimental scRNA-seq datasets. Both methods outperformed PhenoGraph, with Sil of 0.398.

These results were statistically significant in a paired T-test. SOEA was no worse than Kmeans (Sil of 0.429) and outperformed it on 6 datasets. MOEA significantly outperformed Kmeans in 12 datasets. External validity was measured using Normal Mutual Information (NMI) on 6 synthetic datasets. The average NMI values were 0.901 for MOEA and 0.898 for SOEA. These values were on par with Kmeans (0.902) and outperformed PhenoGraph (0.677).

Result 2: Clustering results of SOEA and MOEA were consistent and reproducible. To evaluate the consistency of clustering results, we repeated internal and external validation 30 times and examined the standard deviation of performance metrics. The results of MOEA and SOEA were consistent. In internal validation, MOEA and SOEA had standard deviations of 0.009 and 0.003, respectively. In external validation, the standard deviations of NMI were 0.002 for both MOEA and SOEA, respectively.

Result 3: MOEA and SOEA were stable, and the preprocessing workflows were successful. To evaluate the stability of clustering and the effectiveness of our preprocessing steps, we performed metamorphic testing of 4 algorithms in 30 independent executions. Each metamorphic relation was generated by perturbing individual experimental datasets, and changes in Sil scores were noted. The average Sil scores across 48 datasets of each metamorphic relation were not statistically different from the results of the original datasets.

Result 4: As expected, execution times of MOEA were longer than those of SOEA. To study the running times, we designed three computational experiments. Firstly, we compared MOEA and SOEA running times on a single CPU. Secondly, we used parallel computing on 16, 32, 64, 94, or 128 CPUs, followed by the examination of different CPU combinations. Running times of MOEA decreased by 91% when 32 CPUs with four tasks per node and eight CPUs per task were used.

ACKNOWLEDGEMENT

This work was supported by the Wake Forest Research Fellowship from the URECA Center of Wake Forest University. Computations were performed on the Wake Forest University DEAC Cluster, a centrally managed resource with support provided in part by the University. The authors thank Cody Stevens and Nathan P. Whitener for technical assistance.

REFERENCES

[1] J. Alcalá-Fdez, Alberto Fernández, J. Luengo, J. Derrac, and S. García. 2011. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. *Journal of Multiple-Valued Logic and Soft Computing* (2011).

[2] Dinabandhu Bhandari, C. A. Murthy, and Sankar K. Pal. 1996. Genetic algorithm with elitist model and its convergence. *International Journal of Pattern Recognition and Artificial Intelligence* 10, 06 (Sep 1996), 731–747.

[3] Kalyanmoy Deb and Himanshu Jain. 2014. An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems With Box Constraints. *IEEE Transactions on Evolutionary Computation* 18, 4 (Aug 2014), 577–601.

[4] David B. Fogel and Lawrence J. Fogel. 1996. An introduction to evolutionary programming. In *Artificial Evolution (Lecture Notes in Computer Science)*, Jean-Marc Alliot, Evelyne Lutton, Edmund Ronald, Marc Schoenauer, and Dominique Snyers (Eds.). Springer, Berlin, Heidelberg, 21–33.

[5] Félix-Antoine Fortin, François-Michel De Rainville, Marc-André Gardner, Marc Parizeau, and Christian Gagné. 2012. DEAP: Evolutionary Algorithms Made Easy. *Journal of Machine Learning Research* 13, 70 (2012), 2171–2175.

[6] John H. Holland. 1992. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence.* A Bradford Book, Cambridge, MA, USA.

[7] Himanshu Jain and Kalyanmoy Deb. 2014. An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point Based Nondominated Sorting Approach, Part II: Handling Constraints and Extending to an Adaptive Approach. *IEEE Transactions on Evolutionary Computation* 18, 4 (Aug 2014), 602–622.

[8] Jacob H. Levine, Erin F. Simonds, Sean C. Bendall, Kara L. Davis, El-ad D. Amir, Michelle D. Tadmor, Oren Litvin, Harris G. Fienberg, Astraea Jager, Eli R. Zunder, Rachel Finck, Amanda L. Gedman, Ina Radtke, James R. Downing, Dana Pe'er, and Garry P. Nolan. 2015. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* 162, 1 (Jul 2015), 184–197.

[9] J. MacQueen. 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* 5.1 (Jan 1967), 281–298.

[10] Orit Rozenblatt-Rosen, Michael J. T. Stubbington, Aviv Regev, and Sarah A. Teichmann. 2017. The Human Cell Atlas: from vision to reality. *Nature* 550, 76777677 (Oct 2017), 451–453.

[11] Information Systems and Wake Forest University. 2016. High Performance Computing.