Artificial Neural Network System for Cell Classification using Single Cell RNA Expression

Xin Lin; Jiahui Zhong; Minjie Lyu; Sen Lin; Derin B. Keskin; Guanglan Zhang; Vladimir Brusic; Lou T. Chitkushev

University of Nottingham
UK | CHINA | MALAYSIA

University of Nottingham Ningbo China, 199 Taikang East Road, Ningbo, 315100, Zhejiang, China.

First published 2021

University of Nottingham
UK | CHINA | MALAYSIA

Artificial Neural Network System for Cell Classification using Single Cell RNA Expression

Xin LIN
School of Computer Science
University of Nottingham
Ningbo China
scyxl5@nottingham.edu.cn

Jiahui ZHONG
School of Computer Science
University of Nottingham
Ningbo China
jiahui.zhong@nottingham.edu.cn

Minjie LYU
School of Computer Science
University of Nottingham
Ningbo China
minjie.lyu@nottingham.edu.cn

Sen LIN
School of Computer Science
University of Nottingham
Nottingham UK
psysl7@nottingham.ac.uk

Derin B. KESKIN
Dana-Farber Cancer Institute
Harvard Medical School
Boston USA
Derin_Keskin@dfci.harvard.edu

Guanglan ZHANG
HiLab, Metropolitan College
Boston University
Boston USA
guanglan@bu.edu

Lou T. CHITKUSHEV
HiLab, Metropolitan College
Boston University
Boston USA
LTC@bu.edu

Vladimir BRUSIC
School of Computer Science
University of Nottingham
Ningbo China
vladimir.brusic@nottingham.edu.cn

***Abstract* — We implemented an automated system for single-cell classification using artificial neural networks (ANN). Our system takes single-cell gene expression sparse matrices and trains ANN to classify cell types and subtypes. The assemblies of ANNs predict cell classes by voting. We tested the system in a case study where we trained ANNs with a dataset containing approximately 120,000 single cells and tested the resulting model using an independent data set of 13,000 single cells. The overall accuracy of the 5-class classification was 95%. We trained and tested a total of 100 ANNs in 10 cycles. The prediction system demonstrated excellent reproducibility. The analysis of misclassifications indicated that 2% were likely classification errors, while the remaining 3% were likely due to mislabeled types and subtypes in the test set.**

Keywords — ANN, automation of cell classification, gene expression, PBMC, prediction system, supervised machine learning

I. INTRODUCTION

Single-cell transcriptomics (SCT) examines gene expression profiles in individual cells. It is used for biomarker discovering and studies of heterogeneity of cells involved in biological processes [1]. Bulk RNA sequencing (RNA-seq) provides information on average gene expression across thousands or even millions of cells from the same sample. Bulk RNA-seq methods cannot capture cellular heterogeneity. The profiles of cell subtypes present in the sample remain unknown. Single-cell RNA sequencing (scRNA-seq) produces a more detailed view of RNAseq data because it enables us to assign counts of expressed genes to individual cells [2].

Within the same cell type, RNA expression can be quite different between individuals. Numerous factors, both biological and technical, influence changes in gene expression [3,4]. Biological factors include cell development stages, interactions between cells, and cellular responses to biological or environmental stimuli. In addition to genetic inheritance factors, gene expression changes may be due to the genetic program embedded inside each cell (ontogeny), development stage, activation history, healthy or disease status, age, and other factors [5]. Technical factors include sample processing and storage conditions. For example, the extraction of peripheral blood mononuclear cells (PBMC), freezing, fluorescence-activated cell sorting (FACS) steps, or enrichment methods will influence gene expression relative to the previous sample processing step. The high variability of scRNA-seq data makes cell classification, biomarker discovery, and identification of developmental trajectories challenging tasks [6].

More than 3000 SCT sparse matrices data sets have been generated to date using 10x GemCode Technology [7] and made publicly available. However, the number of captured genes per cell is limited and, in our estimate, representing anywhere between 2 to 30% of genes expressed in most single cells. Thus, gene expression matrices generated by scRNA-seq techniques are sparse and are excellent targets for machine learning and data analytics.

When we train a classification system using instances with known class labels, the learning is called supervised learning. If the instances are unlabeled, the instances are grouped by unsupervised learning. Unsupervised clustering algorithms are used to map items to common classes and discover new, meaningful classes [8].

Current data analytics methods of SCT rely on unsupervised clustering [6]. Unsupervised machine learning methods have a disadvantage that they do not scale up well. Each study requires a combination of manual annotation, and

clustering algorithms that perform well on specific datasets may not perform well on datasets from different studies (lack of generalization) [9]. Machine learning requires that every instance (a single cell) in studied datasets is represented using the same set of features (genes in sparse matrices). It is essential that we have standardized data sets for unsupervised machine learning.

Our group has collected, cleaned, and standardized more than 1,000 human datasets [10] and more than 800 datasets of mouse SCT sparse matrices [11]. It is time-consuming to study such large datasets one-by-one, so we have developed a system for automated classification of SCT data. Here we report a system for automation of ANN training and testing using sparse matrices generated by 10x technology.We also report the prediction of single-cell classes from previously unseen data sets.

Our group previously implemented an ANN to classify PBMC into five major cell types: B cells, dendritic cells (DC), monocytes, natural killer (NK) cells, and T cells [10]. The previous study demonstrated that an ANN can be trained for classification of major PBMC cell types. The performance was decent – 90% accuracy was achieved in 5-class classification. Here, we report an extended analysis, classification of PBMC data using a more extensive training data set. In this study, we focused on a) assessing the generalization properties of ANN classifiers for SCT classification tasks, b) identifying ways to improve generalizable accuracy of ANN classification, and c) exploring reproducibility of machine learning using assemblies of ANNs.

II. Study Design

We defined a standardized string of gene names. It maps gene names from multiple genome assemblies [12] to a common name string. All SCT data used in our study were mapped onto this common list that defined standardized sparse matrix format [10]. Before performing classification and the analysis of classification results, we divided our data into training and testing data sets. The overall design of this study involved four steps: preparation of data sets, building ANN classifier that randomly selects initialization seeds to train multiple ANNs, validation of training, and result analysis (Fig. 1).

III. Methods

A. Data

Data were extracted from three public SCT data sources, GEO database (GEOS data set) [13], Broad Institute database (BroadS1 and BroadS2 data sets) [14], and the 10x company demonstration data [15]. A total of 52 SCT data sets were collected and prepared for the analysis. Each data set has a metadata description. Metadata descriptions provide information about sample collection, processing, and the conditions of experiments. We labeled each data set to reflect their PBMC cell types and subtypes, T cells, B cells, DC, Monocytes, and NK cells. For classification, each data set was labeled with one of the five cell type labels. For analysis, the cell subtype was also used for the assessment of correct classifications and misclassifications. Table I shows the number of data sets used in this study, their cell types and sources. Table II shows the total number of cells by cell types and sources.

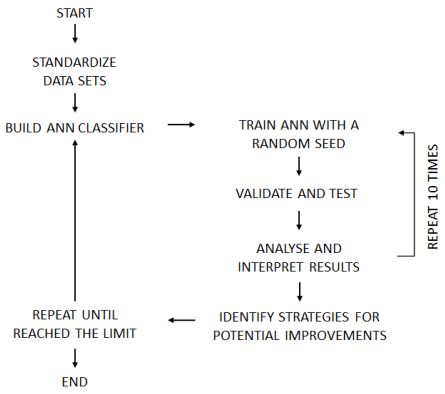

Fig. 1. The overview of this study. In this study we focus mainly on the steps on the right side of the diagram. The improvement strategies will involve the addition of new data sets and cell subtypes, not yet explored in this report.

TABLE I. The number of data sets used in this study. The breakdown of the number of data sets from independent studiesshows, when applicable, splits into partitions representing cell subclasses.

Cell Type	NUMBER OF DATA SETS				
	10xS	GEOS	BroadS1	BroadS2	Total
B cells	1	2	8	4	15
DC	0	0	4	7	11
Monocytes	1	2	8	8	19
NK cells	1	2	4	4	11
T cells	6	20	32	8	66
Total	**9**	**26**	**56**	**31**	**122**

TABLE II. The total number of cells used in this study, by cell types and the sources.

Cell Type	TOTAL NUMBER OF CELLS				
	10xS	GEOS	BroadS1	BroadS2	Total
B cells	9,724	1,786	1,660	1,877	15,047
DC	0	0	142	270	412
Monocytes	1,843	856	1,661	2,007	6,367
NK cells	8,179	618	1,394	842	11,033
T cells	62,649	26,629	8,326	7,151	104,755
Total	**82,395**	**29,889**	**13,183**	**12,147**	**137,614**

B. Quality control

All data used in this study passed our internal quality control. Cells that have 300 or more positive features and 670 or more total counts were selected. We determined the 300 and 670 thresholds empirically from observations of feature count distributions – these thresholds are mostly within the linear part of the S-curve representing the indexed list of counts. The high-end thresholds were not applied in this study (*e.g.* removing cells that have number of features or counts larger than some defined value).

C. Artificial Neural Networks

We used the same architecture, training algorithm, and stopping criteria, as reported earlier [10]. In short, the ANN architecture was 30698-10-5, representing the number of units in input, hidden, and output layers. The MLPClassifier function from Scikit-learn python library was employed to encode a multi-layer perceptron classifier. We used the following parameters: activation: rectified linear unit (ReLU),

solver: adam, alpha: 0.0001, batch size: 200, initial learning rate: 0.001. The default values were used for other parameters.

Each training-testing-assessment run consisted of ten cycles. One ANN was trained in each cycle, using a randomly generated initialization seed. This ensured that the trained ANN models were different in each cycle. The results of each run were analyzed for reproducibility. To assess generalization, we performed ten runs of the system and compared the results of all individual cycles within each run. Reprodycibility between cycles was assessed by the comparison of composite results for each cycle. More details can be found in the IMPLEMENTATION section.

D. Assessment of performance

To compare the results of individual cycles, we calculated the values of Precision (PR), Recall (RE), and F1-value for each individual cell class, and the overall accuracy (ACC) across five classes:

$$PR = \frac{TP}{TP + FP} \qquad RE = \frac{TP}{TP + FN}$$

$$\mathrm{F1} = 2 \times \frac{PR \times RE}{PR + RE}$$

$$ACC = \frac{\sum TP}{\sum TP + \sum FN}$$

where TP, FP, and FN stand for true positives, false positives, and false negatives, respectively. ACC is calculated by using sums of TP and FN across the classes.

IV. IMPLEMENTATION

A. Data management system

Data management for SCT technology is challenging because sparse matrices make Big Data [16]. Because SCT technology is developing rapidly, there is a lack of standardized data sets and of general data standards [17]. In addition, unsupervised machine learning methods, that are methods of choice for SCT data analysis, do not generalize well [18]. Unsupervised methods, therefore, cannot be automatically applied for the analysis of data from multiple studies. The raw data are available in five formats (TXT, CSV, TSV, H5, and MTX). The raw data formats were converted into MTX format (math.nist.gov/MatrixMarket/formats.html) using an in-house software. In a sparse matrix, 95-99% of features are typically zero. In a file with >30,000 rows and >10,000 columns, one CSV file may exceed one Gb. MTX file for sparse matrices stores only non-zero values and their coordinates, thus reducing the memory requirements by approximately an order of magnitude. Files were named so that they inform the user about the cell type, data source, and the cell number in each sparse matrix.

B. ANN prediction system

Our system uses random initialization seeds for ANN to ensure the diversity of initial models. ANN prediction method with only one ANN model may produce a high accuracy model in one study, but it will not necessarily have similar accuracy in other studies. Our system trains multiple ANN models that are not mutually identical to ensure that the assessed accuracy is realistic, and the models generalize well.

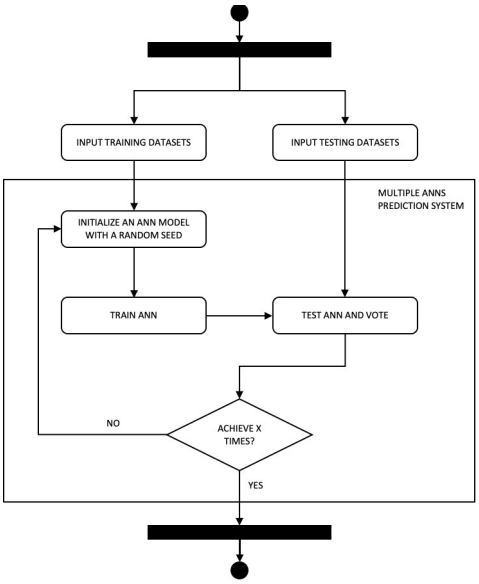

Fig 2. Activity diagram of ANNs prediction system showing multiple iterations within a training cycle. The number of iterations of training-testing can be changed as needed, in our case study it was set to X=10.

We performed a case study to test the system and demonstrate prediction capabilities. Training datasets were from 10x, GEO, and BroadS2, while the testing dataset was BroadS1. Our system repeats training and testing cycles 10 times by default. Each of the 10 ANNs was trained and tested using an identical procedure, so they are fully comparable. The sum of the signals from five output units (representing five PBMC classes) in a single ANN system is always one: $B_o+DC_o+MC_o+NK_o+T_o=1$. The overall prediction score in our system (using 10 models per cycle) is 10. The final classification result of each 10-models cycle is presented as a set of 11 confusion matrices: one matrix for the overall classification result and one matrix for each individual model. The prediction result can be used for performance assessment and misclassification analysis. The activity diagram of our ANNs prediction system is shown in Fig. 2.

C. Case study

In this study, we predicted five main subtypes of peripheral blood mononuclear cells (PBMC). Our case study has extended the PBMC analysis reported in [10]. We performed ten cycles of analysis, where each cycle had ten training-testing iterations. A total of 100 ANNs were trained in ten prediction cycles, and the results of cycles were used to assess the reproducibility and generalizability of our prediction system.

V. RESULTS

A. ANN prediction results

The accuracy of individual ANN models ranged from Acc=0.902 to Acc=0.949. The average value of individual accuracy across the 100 ANN models was 0.939±0.010. The overall accuracy of the 10 cycles, based on voting strategy, was between Acc=0.943 and Acc=0.949. The average of all 10 cycle accuracies was Acc=0.947±0.002. The results for Cycle 1 are shown in Fig. 3. The overall accuracy of classification using voting in 10 iterations is higher than the average accuracy of ANNs in each cycle.

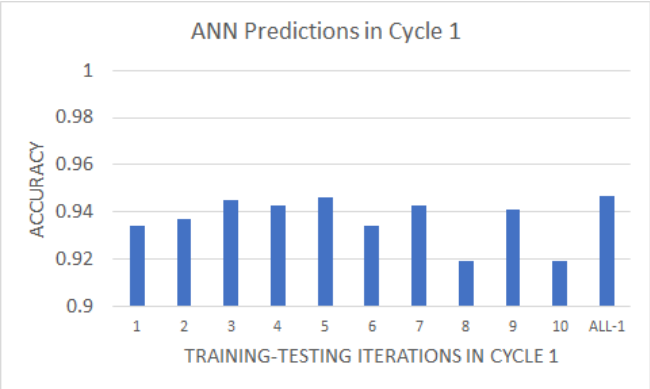

Fig. 3. Ten single ANN prediction results and the overall accuracy based on voting, from Cycle 1 of our case study. Minimum value was $Acc_{1min} = 0.919$, maximum value $Acc_{1max} = 0.946$, average value $Acc_{1ave} = 0.936 \pm 0.001$. The overall accuracy by voting was $Acc_{1vot} = 0.947$.

B. *Voting table*

A snapshot of the voting table is shown in TABLE III. The full voting table has 13,183 rows (see TABLE II, BroadS1). The first row shows the perfect prediction of all output units in all 10 iterations. The second row shows a minor deviation from the perfect score but no misclassification. The third row shows misclassification in some of the individual networks, yet accurate voting. The last two rows show the misclassification of T cells as NK cells. The results of the misclassification analysis are shown in TABLE IV. The true positive rates for predicting B cells, DC, monocytes, NK cells, and T cells were 94.6%, 81%, 96.9%, 83%, and 96.5%, respectively. The largest number of misclassification by percentage was for DC and NK cells. About 17% of DC were misclassified as monocytes. 16.9% of NK cells were misclassified as T cells. 3.5% of T cells were misclassified as NK cells. All other class-class misclassifications were below 2.5%. The poor classification of DC can be explained by the small number of DC instances in the training set (only 270 of 124,431 or 0.22% of the total). Furthermore, DC in the training set were composed of a mix of two subtypes (conventional and plasmacytoid DCs), while the test set did not have subclass annotation. Misclassification of NK cells to T cells and *vice versa* can, at least partially, be explained by the existence of NK-like T cells [19]. Further analysis is needed to resolve classification discrepancies.

TABLE III. AN EXAMPLE OF VOTING TABLE FOR ONE CYCLE OF PREDICTIONS (CYCLE 1). THE INTERPRETATION OF THE RESULTS IS IN THE MAIN TEXT.

BC	DC	MC	NK	TC	PREDICTED	LABELED
0.000	0.000	0.000	0.000	10.000	TC	TC
0.000	0.000	0.098	0.000	9.902	TC	TC
0.000	0.000	0.001	2.286	7.713	TC	TC
0.035	0.012	0.015	5.485	4.453	NK	TC
0.034	0.012	0.015	9.916	0.022	NK	TC

I. CONCLUSIONS AND DISCUSSION

We developed and implemented an ANN-based prediction system that deploys an assembly of ANNs that classify single cell types by assembly voting. Using a test case of PBMC classification, we demonstrated that ANN training, using large-scale data sets generated from different and unrelated single-cell studies (10x technology), are excellent multi-class predictors. The overall accuracy of ANN assembly voting is 95%, and the prediction results were reproducible across all training-testing cycles.

TABLE IV. A REPRESENTATIVE CONFUSION MATRIX USED FOR MISCLASSIFICATION ANALYSIS.

		PREDICTED					
		BC	DC	MC	NK	TC	TOTAL
LABELED	BC	1,570	14	5	45	26	**1,660**
	DC	0	115	24	0	3	**142**
	MC	3	6	1,610	0	42	**1,661**
	NK	1	0	1	1,157	235	**1,394**
	TC	0	0	0	290	8,037	**8,327**
	TOTAL	**1,574**	**135**	**1,640**	**1,492**	**8,343**	**13,184**

A vast majority of correctly classified cells class show excellent agreements with individual predictors. The majority of misclassified cells also have uniform voting profiles. The assembly method shows a small but consistent improvement across all cycles and individual training-testing iterations. The assembly vote matches or exceeds the accuracy of the best individual ANN within a given cycle. In summary, we conclude that the proposed ANN method produces highly reproducible classification results. We estimate that 98% of correct predictions have uniform votes across assembly, and 2% have ambiguous votes. Approximately 80% of misclassified cells have uniform votes. We hypothesize that most of these represent subclasses that are not clearly defined in our data set, such as NK-like CD8+ T cells [19].

Further studies will focus on in-depth analysis of features that characterize misclassified cells and identify their true class. We estimate that, most likely, the true misclassification rate in our system is approximately 2%. We plan to deploy our system for the classification of other cell types with available SCT data sets.

ACKNOWLEDGMENT

This work was supported by Ningbo Service Industry S&T Programme, project code: 2019F1028.

REFERENCES

[1] I. Kanter and T. Kalisky, “Single Cell Transcriptomics: Methods and Applications”, Front. Oncol., vol. 5, 53, 2015.

[2] G. Chen, B. Ning and T. Shi, “Single-cell RNA-seq technologies and related computational data analysis”, Front. Genet., vol. 10, 317, 2019.

[3] A. A. Kolodziejczyk, J. K. Kim, V. Svensson, J. C. Marioni and S. A. Teichmann, “The technology and biology of single-cell RNA sequencing”. Mol. Cell, vol. 58, no. 4, pp.610-620, 2015.

[4] Á. Arzalluz-Luque, G. Devailly, A. Mantsoki and A. Joshi, “Delineating biological and technical variance in single cell expression data”, Int. J. Biochem. Cell Biol, vol. 90, pp.161-166, 2017.

[5] L. M. Lepone, R. N. Donahue, I. Grenga, S. Metenou, J. Richards, C. R. Heery, et al., “Analyses of 123 peripheral human immune cell subsets: defining differences with age and between healthy donors and cancer patients not detected in analysis of standard immune cell types”, J. Circulating Biomark, vol. 5, 5, Jan 1 2016.

[6] D. Lähnemann, D. Köster, J. Szczurek, E. McCarthy, D. J. Hicks, S. C. Robinson, et al., "Eleven grand challenges in single-cell data science", Genome Biol., vol. 21, no, 1, pp.1-35, 2020.

[7] G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, et al., “Massively parallel digital transcriptional profiling of single cells”, Nat. Commun., vol. 8, no. 1, pp. 1-2, 2017.

[8] S. B. Kotsiantis, “Supervised machine learning: a review of classification techniques”, Informatica, vol. 31, pp. 249–268, 2007.

[9] L. Chen, Y. Zhai, Q. He, W. Wang and M. Deng, “Integrating deep supervised, self-supervised and unsupervised learning for single-cell RNA-seq clustering and annotation”, Genes, vol. 11, no. 7, 792, 2020.

[10] R. A. Shaikh, J. Zhong, M. Lyu, S. Lin, D. Keskin, G. L. Zhang, et al., “Classification of five cell types from PBMC samples using single cell

transcriptomics and artificial neural network". In 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2207-2213, IEEE, 2019.

[11] B. Zheng, M. Lyu, S. Lin and V. Brusic, "Tissue of origin classification from single cell mRNA expression by artificial neural networks", In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), this issue, IEEE, 2020.

[12] D. M. Church, V. A. Schneider, T. Graves, K. Auger, F. Cunningham, N. Bouk, H et al., "Modernizing reference genome assemblies", PLoS Biol, vol. 9(7), e1001091, Jul 5 2011.

[13] E. Clough and T. Barrett, "The gene expression omnibus database. In Statistical Genomics", Humana Press, New York, NY, pp. 93-110, 2016.

[14] J. Ding, X. Adiconis, S. K. Simmons, M. S. Kowalczyk, C. C. Hession, N. D. Marjanovic, et al., "Systematic comparison of single-cell and single-nucleus RNA-sequencing methods", Nat. Biotechnol., vol. 38, June, pp. 737–746, 2020.

[15] G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, et al., "Massively parallel digital transcriptional profiling of single cells", Nat. Commun., vol. 8, 14049, 2017.

[16] B. D. Aevermann, M. Novotny, T. Bakken, J. A. Miller, A. D. Diehl, D. Osumi-Sutherland, et al., "Big Data and single cell transcriptomics: implications for ontological representation", Hum. Mol. Genet., vol. 21, R1, pp. R4-R47, 2018.

[17] D. Lähnemann, D. Köster, J. Szczurek, E. McCarthy, D. J. Hicks, S. C. Robinson, et al., "Eleven grand challenges in single-cell data science", Genome Biol., vol. 21, no. 1, pp.1-35, 2020.

[18] W. Hou, Z. Ji, H. Ji and S. C. Hicks, "A systematic evaluation of single-cell RNA-sequencing imputation methods", Genome Biol., vol. 21, 218, 2020.

[19] A. Barbarin, E. Cayssials, F. Jacomet, N. G. Nunez, S. Basbous, L. Lefèvre, et al., "Phenotype of NK-like CD8 (+) T cells with innate features in humans and their relevance in cancer diseases", Front. Immunol., vol. 8, 316, 2017.