ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20191002034614
Original scientific paper

Clustering Single-cell RNA-sequencing Data based on Matching Clusters Structures

Yizhang WANG, You ZHOU, Wie PANG, Yanchun LIANG, Shu WANG

Abstract: Single-cell sequencing technology can generate RNA-sequencing data at the single cell level, and one important single-cell RNA-sequencing data analysis method is to identify their cell types without supervised information. Clustering is an unsupervised approach that can help find new insights into biology especially for exploring the biological functions of specific cell type. However, it is challenging for traditional clustering methods to obtain high-quality cell type recognition results. In this research, we propose a novel Clustering method based on Matching Clusters Structures (MCSC) for identifying cell types among single-cell RNA-sequencing data. Firstly, MCSC obtains two different groups of clustering results from the same K-means algorithm because its initial centroids are randomly selected. Then, for one group, MCSC uses shared nearest neighbour information to calculate a label transition matrix, which denotes label transition probability between any two initial clusters. Each initial cluster may be reassigned if merging results after label transition satisfy a consensus function that maximizes structural matching degree of two different groups of clustering results. In essence, the MCSC may be interpreted as a label training process. We evaluate the proposed MCSC with five commonly used datasets and compare MCSC with several classical and state-of-the-art algorithms. The experimental results show that MCSC outperform other algorithms.

Keywords: clustering; consensus function; single-cell sequencing

1 INTRODUCTION

Single-cell sequencing is a recently developed technique for better understanding cellular heterogeneity [1, 2], and it generates RNA sequencing data that consist of cells with many genes. The single-cell RNA sequencing data analysis has attracted much attention in the field of bioinformatics, especially in identifying the cell types. However, it is a big challenge for analysing single-cell RNA-sequencing data effectively. This is because single-cell RNA-sequencing data has very high dimensions and high level of noise [3], only some dimensions (genes expression levels) differ much, i.e., most attributes may not be helpful for identifying cell types. The simple and easy-understanding way of analysing single-cell RNA-sequencing data is to use clustering algorithms, which are unsupervised learning methods without using class labels. More specifically, clustering algorithms are the methods of grouping data points into multiple clusters with an objective function or a clusters structure hypothesis, such as K-means clustering algorithm [4], density-based spatial clustering algorithm with noise [5], affinity propagation clustering algorithm [6] and spectral clustering algorithm [7]. However, the above-mentioned traditional clustering algorithms cannot work well for analysing single-cell RNA sequence data, because traditional metrics (such as Euclidean distance) are not valid when data points become sparse in high dimensional space. An alternative similarity metric is based on shared nearest neighbour, which is proven to be an effective and robust way of describing relationships between data points in high dimensional space [8]. Concretely, the shared nearest neighbour is the intersection of neighbouring points of a pair of data points.

There exist some methods which can group data into different clusters based on shared nearest neighbour. Guha et al. proposed a robust clustering algorithm for categorical attributes based on the number of neighbouring points to clustering categorical data [9]. Jarvis et al. built a near neighbour list of every data point so as to compute similarities [10]. Ertoz et al. proposed an improved density-based clustering algorithm based on shared nearest neighbour to identify clusters of varying densities and shapes [11]. Based on previous successful applications, shared nearest neighbour is proven to be capable of better revealing the relationships among data points in high-dimensional space [12].

Based on the advantages of shared nearest neighbour similarity, we propose a novel clustering algorithm called Matching Clusters Structures-based Clustering algorithm (MCSC). Five commonly used public real-world datasets are used to evaluate the proposed MCSC and we compare it with four classical methods (Spectral clustering algorithm, K-means clustering algorithm, principal component analysis [13], and t-distributed stochastic neighbour embedding [14]) as well as two state-of-the-art methods (they will be described in Section 2).

The rest of this paper is organized as follows. In Section 2, we review related work. In Section 3, we present the details of our proposed MCSC method. In Section 4, we report the experimental results with discussions. In Section 5, we conclude the paper and propose future work.

2 RELATED WORKS

Clustering is an important approach to identify single cell among RNA sequence data, which has attracted great attention of many researchers.

To the best of our knowledge, many algorithms are proposed to identify cell types and help find new insights into biology. To name a few, Jiang et al. designed a similarity measure based on differentiability correlation between cell pair and then cooperated with hierarchical clustering to form a variance analysis-based clustering algorithm, which can find the true number of clusters automatically and identify cell types efficiently [15]. Wolf et al. developed a scalable tool kit to clustering single cell RNA sequencing data [16]. Kiselev et al. proposed a single-cell consensus clustering method, which is a useful tool for unsupervised clustering [17]. Nikolenko et al. introduced a novel algorithm based on hamming graphs and bayesian sub-clustering for error correction in single-cell sequencing data [18]. Aibar et al. developed a computational method for simultaneous gene regulatory network reconstruction and cell-state identification from single-cell RNA sequencing data [19]. Seyoung Park and Hongyu Zhaouse multiple doubly stochastic similarity matrices to learn a similarity called MultiPle similarity Sparse Spectral Clustering algorithm (MPSSC) [20]. Xu et

al. proposed a clustering algorithm incorporating a shared nearest neighbour graph and quasi-clique recognition methods used to identify cell types from single-cell transcriptomes [21]. Wang et al. proposed a Single-cell Interpretation method via Multi-kernel LeaRning (SIMLR), which improves the visualization and interpretability of single cell RNA sequencing data [22]. Both SIMLR proposed by Wang et al. and MPSSC proposed by Park et al. are based on metric learning. We select SIMLR and MPSSC as benchmarking models because they are well-recognised algorithms.

Previous methods did not consider an unsupervised learning method, which combines different initial clusters into a unified cluster based on their structures matching degree. In this research, we propose a novel clustering method based on matching clusters structures, namely MCSC. It combines multiple different grouping results from the same dataset with an aim to produce superior results.

3 CLUSTERING BASED ON MATCHING CLUSTERS STRUCTURES (MCSC)

In this section, we present the proposed MCSC, which uses a consensus function based on matching clusters structures to decide if two initial clusters are merged into one or not. MCSC first uses K-means clustering algorithm to generate initial clusters because it can obtain stable clustering results by using neighbour information. For the two groups of results of K-means: R^i and R^j, we design a novel consensus function based on shared nearest neighbour to train the results of K-means. Based on the shared nearest neighbour information between different initial clusters, one cluster may be merged into the other. We give a consensus function to determine whether the merging process is reasonable or not.

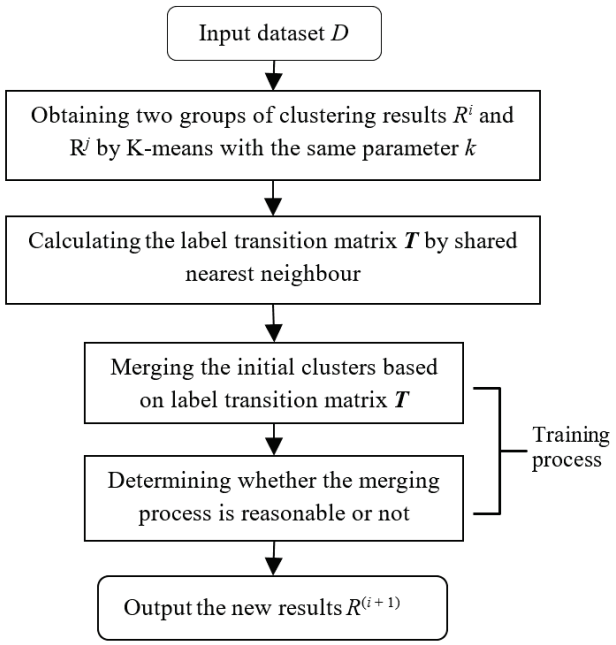


Figure 1 Flow chart of the proposed algorithm MCSC

Finally, the categories of some original initial clusters will change, and we take the final results as output. In order to illustrate our algorithm more intuitively, the flow chart of the proposed algorithm MCSC is shown in Fig. 1.

We first introduce two basic tools: K-means clustering algorithm and a popular external evaluation criterion called Normalized Mutual Information (NMI) in Sections 3.1 and 3.2. The details of our algorithm are described in Section 3.3. The time complexity of the proposed MCSC is given in Section 3.4.

3.1 K-means

We choose the well-known K-means clustering algorithm as the initial clusters' generation methods [23]. The objective function of K-means is defined as follows:

$$\min \sum_{i=1}^{k} \sum_{x_i \in C_i} \|x_i, c_i\| \tag{1}$$

where x_i is a data points in a dataset $D = (x_1, x_2, \ldots, x_N)^T$, c_i is the centroids in an initial cluster C_i and k denotes the number of centroids. Note that k initial centroids are randomly selected so that the clustering results may be different even though parameter k is fixed as shown in Tab. 1, F-measure is chosen as the evaluation metric. F-measure is a commonly used evaluation metric, and for a pair of points (x_i, x_j), they are represented as TP if they have the same label and the same cluster. They are represented as FP if they have different labels but are grouped into the same cluster. They are represented as FN if they have the same label but are grouped into different clusters.

$$Precision = \frac{\#TP}{\#TP + \#FP} \tag{2}$$

$$Recall = \frac{\#TP}{\#TP + \#FN} \tag{3}$$

$$F\text{-}measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

In the above equations, $\#TP$ represents the number of data points belonging to TP, $\#FP$ denotes the number of data points belonging to FP and $\#FN$ denotes the number of data points belonging to FN.

Table 1 Five times clustering results obtained by K-means (k = 3) on a synthetic dataset

Times	1	2	3	4	5
F-measure	0.9766	0.7326	0.7326	0.7326	0.9766

In brief, we use R^i to denote the i-th time results of K-means when we fix the parameter k, perhaps $R^i \neq R^{i+1}$. Many methods are proposed to improve K-means, such as automatically selecting k or centroids.

3.2 Normalized Mutual Information (NMI)

In the proposed MCSC algorithm, we take Normalized Mutual Information (NMI) as a consensus function to measure clusters structures similarity of any two initial clusters [24]. NMI is a popular external evaluation criterion for cluster quality. For ground-truth A and a group of clustering result B of a dataset D, the unique value in A is defined as a vector $\boldsymbol{X}$ and the unique value in B is defined as a vector $\boldsymbol{Y}$. Thus, the NMI value of two vectors A and B is defined as follows:

$$NMI(A,B) = 2\frac{I(\boldsymbol{X};\boldsymbol{Y})}{H(\boldsymbol{X})+H(\boldsymbol{Y})} \tag{5}$$

$$I(\boldsymbol{X};\boldsymbol{Y}) = \sum_{y\in\boldsymbol{Y}}\sum_{x\in\boldsymbol{X}} \log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{6}$$

$$H(\boldsymbol{X}) = -\sum_{1}^{n} p(x_i)\log_2 p(x_i) \tag{7}$$

$$H(\boldsymbol{Y}) = -\sum_{1}^{n} p(y_i)\log_2 p(y_i) \tag{8}$$

where $p(x)$ denotes the probability of x in A, $p(y)$ denotes the probability of y in B, and $p(x, y)$ denotes the joint distribution probability of x and y. Usually, we use ground-truth and clustering results to compute the NMI value, whose range is between [0, 1]. If the NMI value is closer to 1, the quality of clustering result is higher.

3.3 Clustering Based on Matching Clusters Structures (MCSC)

In this section, we present the steps of the proposed MCSC.

Step1: we obtain two groups of clustering results R^i and R^j by K-means with the same parameter k.

We run K-means twice and keep the parameters unchanged each time. Then, we obtain two groups of results: R^i and R^j which contain the categorization information. MCSC first deals with R^i and then uses the R^j to train the R^i with a consensus function.

Step2: calculate the label transition matrix $\boldsymbol{T}$.

Shared nearest neighbour can effectively represent the structures of high-dimensional data [25]. In MCSC, we use it to find the relationship between two initial clusters. We use A_{ij} to denote the j-th nearest neighbour of data point x_i .The similarity between data point x_i and x_j is defined as follows:

$$S(x_i, y_j) = \sum(knum - m + 1)(knum - n + 1) \tag{9}$$

where *knum* denotes the number of shared nearest neighbour of data points x_i and x_j, m and n denotes the position of shared nearest neighbours in data points x_i and x_j nearest neighbour list (the i-th and j-th rows of $\boldsymbol{A}$), respectively. When $S(x_i, x_j) > \theta$, data points x_i and x_j be connected and they are very likely to be merged into one cluster. The parameter θ is user-defined. We use $con(x_i, x_j)$ to denote the connected state of data points x_i and x_j.

For all the initial clusters C_i in R^i, we give a strong hypothesis that a bigger initial cluster is more likely to be a major part of a natural cluster. Thus, the clusters with small number of data points may be merged into bigger one. MCSC decides which initial clusters are micro-clusters (*mc*) based on their number of data points. The micro-clusters (*mc*) are defined as follows:

$$mc = \left\{ C_i \middle| |C_i| < \frac{1}{k}\sum_{i}^{k}|C_i| \right\}, C_i \in R^i \tag{10}$$

where $|C_i|$ denotes the number of data points in the initial cluster C_i. We refer to the remaining initial clusters as core clusters (*cc*). Then we propose a transition matrix to denote the shared nearest neighbour information between micro-clusters (*mc*) and core clusters (*cc*). The transition matrix $\boldsymbol{T}$ between *mc* and *cc* is defined as follows:

$$\boldsymbol{T}(C_i, C_j) = \frac{\#con(x_i, x_j)}{|C_i||C_j|} \tag{11}$$

Where $\#\ con(x_i, x_j)$ denotes the number of the pair of data points that can be connected, $x_i \in C_i$, $x_j \in C_j$, $C_i \in mc$, $C_j \in cc$, $R^i = cc \cup mc$. T_{ij} describes the probability that the micro-cluster C_i is merged into the core cluster C_j. Obviously, for each micro-cluster C_i, we may merge it into certain core cluster C_j with maximum probability. To assess the rationality of the merging process, we employ a well-known metric called normalized mutual information (NMI), which is often used to measure clusters structures similarity of any two initial clusters. If the NMI value of the results after the merging process and the other group of clustering result of K-means (R^j) increases, the process is successful. Then the proposed method MCSC will continue to assign the remaining micro-clusters. The whole label training details are given in the next step.

Step3: training the results of K-means by consensus function.

For two groups of results R^i and R^j generated by K-means, we use a well-established cluster validity index normalized mutual information (NMI) to construct a consensus function:

$$max\ NMI(R^i, R^j) \tag{12}$$

Obviously, the unique value in R^i is defined as a vector $\boldsymbol{X}$ and the unique value in R^j is defined as a vector $\boldsymbol{Y}$. We take a group of clustering result as ground-truth in order to train the other group of results. When all the data points of C_i are merged into C_j for R^i, if $NMI(R^i, R^j)$increases, the merging process is reasonable.

In short, each micro-cluster C_i is merged into the most suitable core cluster, then MCSC will obtain the results that satisfy Eq. (12). The solution procedure of Eq. (12) is presented in Algorithm 1.

Algorithm 1 The solution procedure of Eq. (12)

Input: R^i and R^j: two groups of clustering results obtained by K-means, label transition matrix T.
Output: $R^{(i+1)}$
$k \leftarrow$ the number of micro-clusters in R^i;
For i=1→k**do**
  T_{im} ←the non-zero maximum value in i-th rows of T;
  If NMI(R^i,R^j) increases then
    Ripoints ← the corresponding data points of C_m in R^i
    goallabel← the label with the largest number in *Ripoints*;
    merge all the data points of $C_i \in R^i \cup C_m \in R^i$into $C_{goallabel}$;
  else
Restore the labels of C_i and C_m;
  end
end
output the new results $R^{(i+1)}$

MCSC has three parameters k, θ and *knum*. Parameter k denotes the number of centroids, parameter θ denotes the cut-off value of distances among data points, then decides which two data points are regarded as $con(x_i, x_j)$. Parameter

knum denotes the number of nearest neighbours. In implementation, we set $\theta = 0$ and *knum* = 5, k is depending on datasets.

After the above training process, we obtain the final results $R^{(i+1)}$. In implementation, if R^i is closer to ground-truth, the final results will be better.

3.3 Complexity Analysis

The MCSC algorithm is based on k-means, whose time complexity is $O(n\log n)$. In the process of calculating the label transition matrix, the time complexity is $O(n^2)$. Then in the label training process, the time complexity is $O(n)$. Thus, time complexity of the proposed MCSC is $O(n^2)$. In terms of time complexity, the MCSC algorithm is not higher than other benchmarking models.

4 EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we use five single-cell RNA datasets to evaluate the performance of our proposed method and analyse the results.

4.1 Datasets

The datasets can be downloaded online[1], the sources of data are in the brackets, i.e., Deng, Treutlein, Pollen, Tasic and Buettner datasets are from reference [26-30], respectively. The dataset features are shown in Tab. 2, where the \#cell denotes the number of samples, \#genes denotes dimensions of datasets and \#cell types denotes the labels.

Table 2 Datasets features

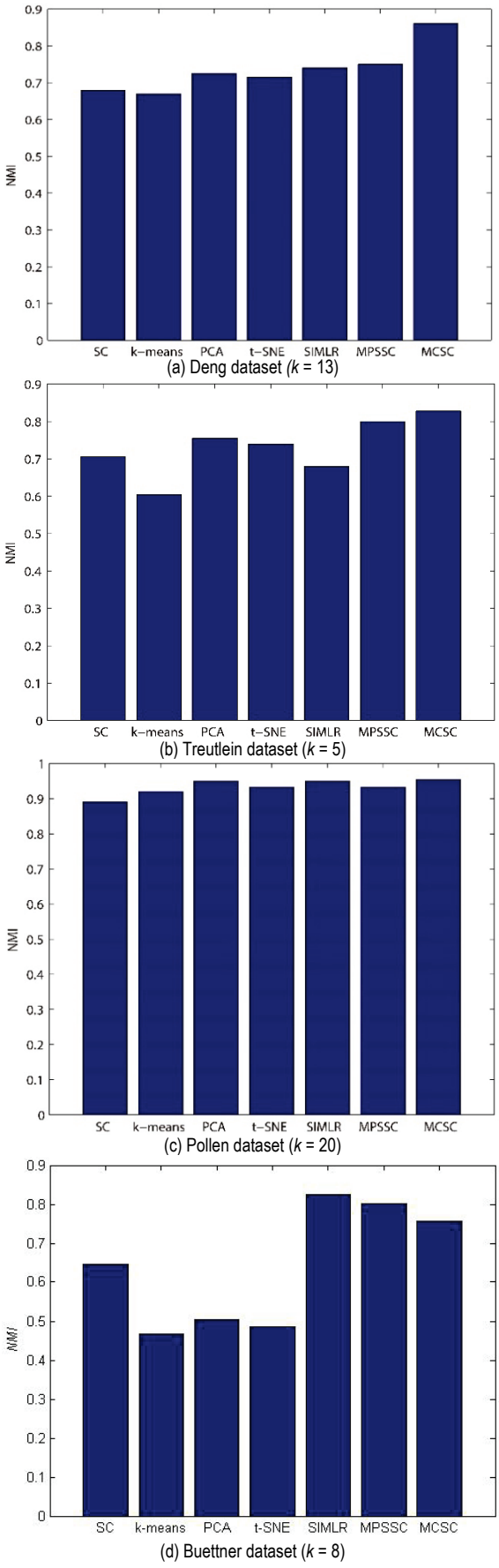
Datasets	#cell	#genes	#cell types
Deng [26]	135	12548	7
Treutlein [27]	80	9352	5
Pollen [28]	249	14805	11
Tasic [29]	1727	5832	49
Buettner [30]	182	8989	3

4.2 Results

We compare MCSC with four classical methods (Spectral clustering algorithm (SC), K-means clustering algorithm, Principal Component Analysis (PCA) [13] and t-distributed stochastic neighbour embedding (t-SNE) [14]) and two state-of-the-art methods (multiple similarity sparse spectral clustering algorithm (MPSSC) and single-cell interpretation method via multi-kernel learning (SIMLR)) in Fig. 2.

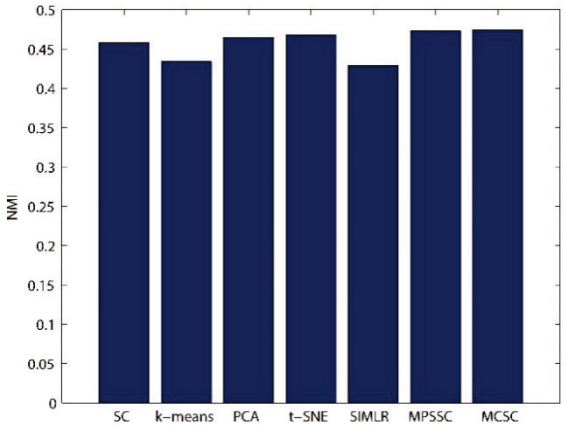
We take NMI as evaluation index. The codes of SC, K-means, PCA, t-SNE, SIMLR and MPSSC can be downloaded online [2]. We use MATLAB R2014a to implement our algorithm and present the best results among 100 times trials. Note that we use the raw data without pre-processing.

In the meantime, the running time of all algorithms including four classical methods (Spectral clustering (SC), K-means, Principal Component Analysis (PCA) and *t*-distributed stochastic neighbour embedding (t-SNE) and two state-of-the-art methods (SIMLR and MPSSC) are shown in Fig. 3.



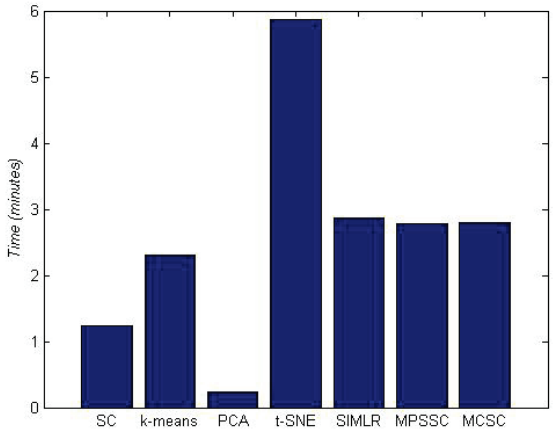
[1] https://github.com/ishspsy/project/tree/master/MPSSC

[2] https://github.com/ishspsy/project/tree/master/MPSSC

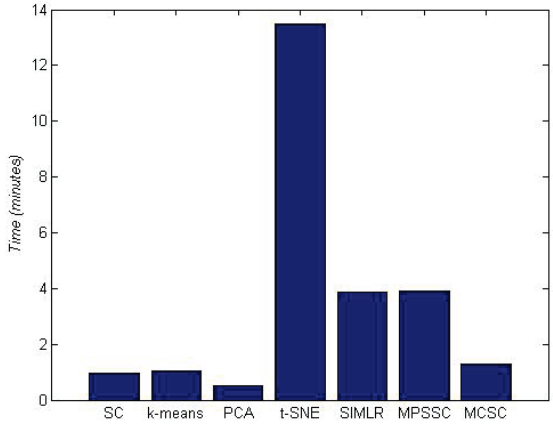


(e) Tasic dataset (k = 100)

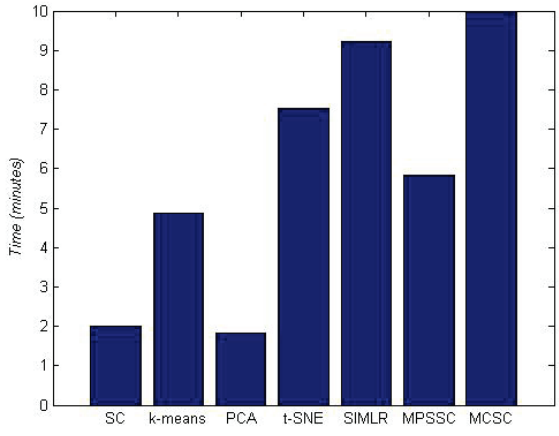
Figure 2 Evaluation of the seven clustering methods by NMI. Values of the parameter k are in the brackets.



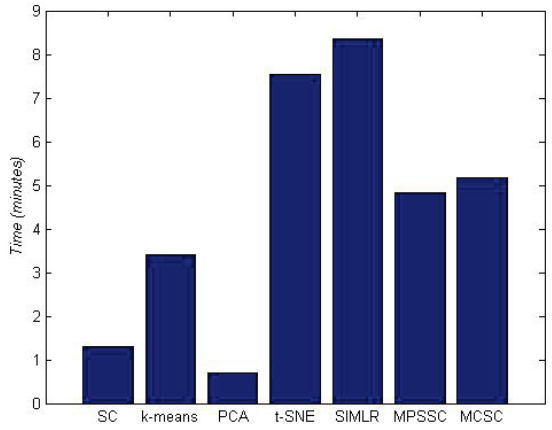
(a) Deng dataset



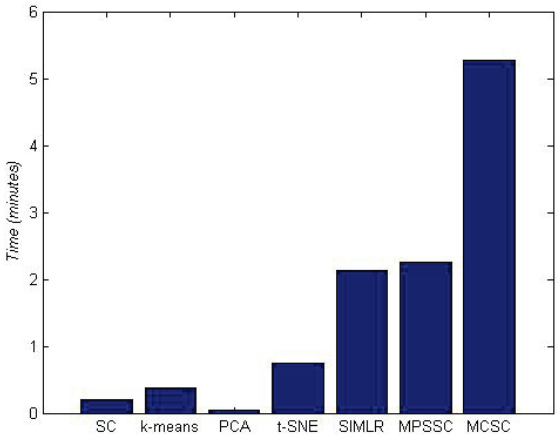
(b) Treutlein dataset



(c) Pollen dataset



(d) Buettner dataset



(e) Tasic dataset

Figure 3 Running time of the seven clustering methods

The running time of the first four datasets is measured by seconds and the last is measured by minutes. The values of parameter k are selected as before.

4.3 Discussion

We propose a novel clustering method based on matching clusters structures, namely MCSC. MCSC can improve the results of clustering algorithms by a label training process. Different initial clusters generation method may have a great impact on the results. We select K-means as the initial clusters generation method, which will obtain unsatisfying results when the structure of dataset is not convex.

The centroids selection of K-means is random, so we need to experiment many times. Actually in 100 experimental results, several results are better than other algorithms. In implementation, we fix the parameters θ and *knum*, users only need to adjust the parameter k. The k value is usually larger than the number of classes.

Table 3 100 times results (NMI) of MCSC

Datasets	Mean value	Maximum value	Variance
Deng [26]	0.7602	0.8666	0.0027
Treutlein [27]	0.6860	0.8286	0.0059
Pollen [28]	0.9183	0.9534	0.0006
Tasic [29]	0.4455	0.4746	0.0005
Buettner [30]	0.5846	0.7594	0.0038

As shown in Fig. 2 and Tab. 3, the maximum values of MCSC are better than other algorithms. For Deng and Treutlein datasets, our algorithm has obvious advantages

because the structure of them is suitable for K-means algorithm. The advantage is slightly less for Pollen and Tasic datasets. For Buettner dataset, its structure is non-convex and K-means easily obtains suboptimal results, so the NMI value of MCSC is lower than SIMLR and MPSSC.

From an algorithm runtime point of view, all the algorithms were run on the same device and software. The most advanced algorithms have no advantage in running time especially for MCSC. SC, K-means and PCA take the least time for all the five datasets, while SIMLR, MPSSC and MCSC take more time because they are based on basic algorithms. For all datasets, the running time of MCSC is close to t-SNE, SIMLR and MPSSC.

Overall, the proposed method MCSC obtains better results and requires close time in most cases.

5 CONCLUSION AND FUTURE WORK

In this research, we propose a novel clustering results improvement method based on matching clusters structures without true labels. However, the performance of MCSC depends on the algorithm that generates the initial clusters. If the structure of some datasets is non-convex, MCSC may obtain unsatisfied results because it uses K-means to generate initial clusters. In our future work, we plan to further improve the label training process and choose more appropriate clustering algorithms to obtain initial clusters.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (Grants Nos. 61772227, 61572227), the Science & Technology Development Foundation of Jilin Province (Grants No. 20180201045GX) and the Science Foundation of Education Department of Guangdong Province (Grants Nos. 2017KQNCX251, 2018XJCQSQ026) and the Social Science Foundation of Education Department of Jilin Province (Grants No. JJKH20181315SK).

6 REFERENCES

[1] Pelkmans, L. (2012). Using Cell-to-Cell Variability--A New Era in Molecular Biology. *Science, 336*(6080), 425-426. https://doi.org/10.1126/science.1222161

[2] Brennecke, P., Anders, S., Kim, J. K., et al. (2014). Corrigendum: Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods, 11*(2), 210-210. https://doi.org/10.1038/nmeth0214-210b

[3] Bacher, R. & Kendziorski, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biology, 17*(1), 63. https://doi.org/10.1186/s13059-016-0927-y

[4] Wagstaff, K., Cardie, C., Rogers, S., et al. (2001). Constrained K-means Clustering with Background Knowledge. *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*.

[5] Ester, M., Kriegel, H. P., Sander, J., et al. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *International Conference on Knowledge Discovery & Data Mining*.

[6] Frey, B. J. & Dueck, D. (2007). Clustering by Passing Messages Between Data Points. *Science, 315*(5814), 972-976. https://doi.org/10.1126/science.1136800

[7] Luxburg, U. V. (2007). A tutorial on spectral clustering. *Statistics and Computing, 17*(4),395-416. https://doi.org/10.1007/s11222-007-9033-z

[8] Houle, M. E., Kriegel, H. P., Peer Kröger, et al. (2010). Can Shared-Neighbor Distances Defeat the Curse of Dimensionality? *Scientific and Statistical Database Management, 22nd International Conference*, SSDBM 2010, Heidelberg, Germany, June 30 – July.

[9] Guha, S. (2000). ROCK: A Robust Clustering Algorithm for Categorical Attributes. *Information Systems, 25*. https://doi.org/10.1016/S0306-4379(00)00022-3

[10] Jarvis, R. A. & Patrick, E. A. (1973). Clustering Using a Similarity Measure Based on Shared Near Neighbors. *IEEE Transactions on Computers, C-22*(11), 1025-1034. https://doi.org/10.1109/T-C.1973.223640

[11] Levent, E., Steinbach, M., & Kumar, V. (2003). Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data. *Siam International Conference on Data Mining*. https://doi.org/10.1137/1.9781611972733.5

[12] Yin, J., Fan, X., Chen, Y., et al. (2005). High-Dimensional Shared Nearest Neighbor Clustering Algorithm. *Fuzzy Systems and Knowledge Discovery*. Springer Berlin Heidelberg. https://doi.org/10.1007/11540007_60

[13] Jolliffe, I. T. (2002). Pincipal Component Analysis. *Journal of Marketing Research, 87*(100), 513.

[14] Laurens, V. D. M. (2014). Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research, 15*(1), 3221-3245.

[15] Jiang, H., Sohn, L. L., Huang, H., et al. (2018). Single cell clustering based on cell-pair differentiability correlation and variance analysis. *Bioinformatics, 34*(21), 3684-3694. https://doi.org/10.1093/bioinformatics/bty390

[16] Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome biology, 19*(1), 15. https://doi.org/10.1186/s13059-017-1382-0

[17] V. Y., Kirschner, K., Schaub, M. T., et al. (2017). SC3: consensus clustering of single-cell RNA-seq data. *Nature methods, 14*(5), 483. https://doi.org/10.1038/nmeth.4236

[18] Nikolenko, S. I., Korobeynikov, A. I., & Alekseyev, M. A. (2013). Bayes Hammer: Bayesian clustering for error correction in single-cell sequencing. *BMC genomics. BioMed Central, 14*(1), 7. https://doi.org/10.1186/1471-2164-14-S1-S7

[19] Aibar, S., González-Blas, C. B., Moerman, T., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. *Nature methods, 14*(11), 1083. https://doi.org/10.1038/nmeth.4463

[20] Park, S., Zhao, H., & Birol, I. (2018). Spectral clustering based on learning similarity matrix. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bty050

[21] Xu, C. & Su, Z. (2015) Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics, 31*(12), 1974-1980. https://doi.org/10.1093/bioinformatics/btv088

[22] Wang, B., Zhu, J., Pierson, E., et al. (2017). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nature Methods, 14*(4), 414-416. https://doi.org/10.1038/nmeth.4207

[23] Huang, J., Ng, M., Rong, H., et al. (2005). Automated Variable Weighting in k-Means Type Clustering. *IEEE Transactions on Pattern Analysis & Machine Intelligence, 27*(5), 657-668. https://doi.org/10.1109/TPAMI.2005.95

[24] Pablo, A., Estévez, Member, S., et al. (2009). Normalized Mutual Information Feature Selection. *IEEE Transactions on Neural Networks, 20*(2), 189-201. https://doi.org/10.1109/TNN.2008.2005601

[25] Kumarpatidar, A., Agrawal, J., & Mishra, N. (2013). Analysis of Different Similarity Measure Functions and Their Impacts on Shared Nearest Neighbor Clustering

Approach. *International Journal of Computer Applications, 40*(16), 1-5. https://doi.org/10.5120/5061-7221
[26] Deng, Q., Ramskold, D., Reinius, B., et al. (2014). Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells. *Science, 343*(6167), 193-196. https://doi.org/10.1126/science.1245316
[27] Treutlein, B., Brownfield, D. G., Wu, A. R., et al. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature, 509*(7500), 371-375. https://doi.org/10.1038/nature13173
[28] (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signalling pathways in developing cerebral cortex. *Nature Biotechnology, 32*(10), 1053-1058. https://doi.org/10.1038/nbt.2967
[29] Tasic, B., Menon, V., Nguyen, T. N., et al. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature Neuroscience.* https://doi.org/10.1038/nn.4216
[30] Buettne, F., Natarajan, K. N., Casale, F. P., et al. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology, 33*(2), 155-160. https://doi.org/10.1038/nbt.3102

Contact information:

Yizhang WANG, PhD student
College of Computer Science and Technology, Jilin University,
Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, China
2699 Qianjin Street, Changchun, 130012, China
wyzhang_new@sina.com

You ZHOU, Prof. PhD
College of Computer Science and Technology, Jilin University,
Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, China
2699 Qianjin Street, Changchun, 130012, China
zyou@jlu.edu.cn

Wei PANG, PhD
The School of Natural and Computing Sciences, University of Aberdeen, Aberdeen, UK
pang.wei@abdn.ac.uk

Yanchun LIANG, Prof. PhD
(Corresponding author)
College of Computer Science and Technology, Jilin University,
College of Computer Science, Zhuhai College of Jilin University,
2699 Qianjin Street, Changchun, 130012, China
ycliang@jlu.edu.cn

Shu WANG, PhD
(Corresponding author)
College of Computer Science Zhuhai College of Jilin University
Zhuhai, 519041, China
wangshujlu@163.com