Sensors and Materials, Vol. 34, No. 4 (2022) 1351–1365
MYU Tokyo

S & M 2896

Comparative Study of Multiple Fitting Regression and Bayes and Probabilistic Support Vector Machine Methods in Classification of Single-cell RNA Data

Huoyou Li,[1*] Yiran Wang,[2] Jianjian Yan,[3] Guoli Ji,[2]
Hsien-Wei Tseng,[1**] and Chun-Chi Chen[4]

[1]School of Mathematics and Information Engineering, Longyan University, Fujian 364012, China
[2]Department of Automation, Xiamen University, Fujian 361005, China
[3]School of Software & Internet of Things Engineering, Jiangxi University of Finance and Economics, Jiangxi 330013, China
[4]School of Life Sciences, Longyan University, Fujian 364012, China

(Received July 1, 2021; accepted January 17, 2022)

Keywords: single-cell RNA, PSVM, MFRB, machine learning, data mining

With the development of single-cell RNA sequencing technology, it is very important and valuable to supplement and improve the mining algorithm of single-cell RNA data to understand the heterogeneity of single-cell RNA and the precise mechanism of the prevention and treatment of diseases. Machine learning and data mining are the preferred technologies for processing large amounts of data. The multiple fitting regression and Bayes (MFRB) method is a new method that combines multiple fitting regression (MFR) methods and Bayesian decision-making in machine learning. The probabilistic support vector machine (PSVM) method is excellent for data classification and has been widely used and verified. In this study, these two classification methods were used to detect large-scale single-cell RNA data and small-sample unbalanced single-cell RNA data, respectively. The performances of the two algorithms were determined and their classification effects were discussed. A random walking preprocessing algorithm is also used to improve the distribution characteristics of low-quality data. The results show that the two algorithms have good results only for large-scale single-cell RNA data; for small-sample unbalanced data sets, neither of the algorithms effectively classified single-cell RNA data.

1. Introduction

In recent years, with the progress in the application of preimplantation genetic diagnosis, preimplantation genetic screening (PGS),(1) and tumor target cell determination and treatment,(2) especially circulating tumor cell assessment and detection,(3) single-cell RNA sequencing technology has attracted increasing attention and development in various fields. Starting from Matioli electrophoresis hemoglobin separation technology,(4) Shi, Song, Geng, Wu, and Guan. invented patch-clamp technology(5) and Dalian Institute of Chemical Physics proposed the concept of the microfluidic chip,(6) both of which are powerful tools for single-cell RNA

*Corresponding author: e-mail: huoyouli@126.com
**Corresponding author: e-mail: hsienwei.tseng@gmail.com
https://doi.org/10.18494/SAM3524

ISSN 0914-4935
https://myukk.org/

sequencing technology research. After that, Bendall's flow cytometry mass spectrometry[7] and Liu's capillary zone electrophoresis tandem method[8] provided a more direct and accurate analysis basis for single-cell RNA sequencing technology.[9]

With the development of single-cell RNA sequencing technology from the first generation to the fourth generation, single-cell RNA sequencing technology continues to improve and complement.[9] Although sequencing technology has made considerable progress in increasing the reading length,[10] expanding flux,[11] increasing depth,[12] increasing speed, and reducing cost,[13] it still needs to be improved in terms of the reading length and precision.[14] Moreover, the heterogeneity[15] and variability of single cells[16] increase the computing power required in the sequencing process and have also brought more challenges to the subsequent downstream analysis. Chitsaz and Yee-Greenbaum proposed the SPAdes algorithm to realize the early diagnosis of cancer cells.[17] Miyatake used the Exome Hidden Markov Model with non-negative least squares regression and other algorithms to achieve data denoising in the population sample mode.[18] Vinga and Almeida used Renyi continuous entropy and other methods to measure the complexity of information and to obtain a complete classification.[19] Koslicki used the RaceID algorithm to search for rare types in mixed single-cell RNA data.[20] A recent trend has been to combine machine learning algorithms to optimize single-cell RNA data detection tools. Classification is an important research topic in single-cell RNA data detection techniques. The classification of single-cell RNA data can not only predict unknown cell types[21] but also identify abnormal cell types,[22] preliminarily screen out subtypes of cell types,[23] and detect low-quality cells.[24] Although the existing single-cell identification methods can be of great guiding significance for follow-up work, owing to the heterogeneity or batch effect of the single-cell RNA data itself, the follow-up research has been greatly limited.

Using the common machine learning algorithms to fully test and explain the classification performance of single-cell RNA data, we take single-cell RNA data as an example to conduct an in-depth exploration of the application of algorithms. Machine learning methods, namely, the multiple fitting regression and Bayes (MFRB) and probabilistic support vector machine (PSVM) methods and intelligent manufacturing, realize design process, manufacturing process, and production equipment intelligence through intelligent perception, human–computer interaction, decision-making, and execution technology. In the equipment employing the MFRB and PSVM methods, various intelligent sensors such as robotic arms, laser detectors, and vision cameras provide effective data support for the decision-making and execution of intelligent manufacturing. These algorithms and analyses are all applied to sensors and sensor fusion technologies, and once these algorithms are widely applied in electronic products such as mobile phones, the consumer experience will be greatly improved. With the increase in user requirements for device detection environments, the tasks assigned to sensors have become increasingly complex, and these machine learning algorithms are needed to solve corresponding problems. Consequently, in this paper, the support vector machine (SVM) method, which is widely used in machine learning, was used to conduct a comparative study with the MFRB[25] method, which has already been used for the classification of cancer cells. Two types of single-cell RNA data are taken as examples: large-scale single-cell RNA data sets and small-sample unbalanced single-cell RNA data sets.[26] The expressions of the two classification algorithms

for these two types of data are analyzed, and the distribution characteristics of single-cell RNA data are clarified in detail.

2. Methods

2.1 PSVM method

The SVM output is binary, which ignores the relative confidence in the classification result. To address this shortcoming, some researchers have modified the SVM to provide a probabilistic output. The most popular probabilistic SVM was proposed by Platt,[27] who adopted a sigmoid function to transform the SVM output into a posterior probability output.

Suppose $N+$ and $N-$ are the numbers of positive ($y = +1$) and negative ($y = -1$) samples, respectively, in a data set D. The probability output of the PSVM is

$$P\left(y=1\middle|f\left(\boldsymbol{x}\right)\right)=\frac{1}{1+\exp\left(-Af\left(\boldsymbol{x}\right)+B\right)}, \tag{1}$$

where $f(\boldsymbol{x})$ is the SVM output given by Eq. (2), and the parameters A and B are obtained from minimizing the negative log-likelihood of the data set D:

$$\min F\left(A,B\right)=\min\left(-\sum_{i=1}^{n}t_i\log\left(p_i\right)+\left(1-t_i\right)\log\left(1-p_i\right)\right), \tag{2}$$

where

$$t_i=\begin{cases}\dfrac{N_+ +1}{N_+ +2}, \text{ if } y_i=1,\\ \dfrac{1}{N_- +2}, \text{ if } y_i=-1.\end{cases} \tag{3}$$

$A < 0$ ensures the monotonicity of Eq. (3). From the geometrical theory of the SVM, we find that the SVM output is proportional to the distance from the chosen hyperplane. Therefore, the probability output of the PSVM means that if the distance from the sample to the chosen hyperplane is large, then its probability of belonging to one class is higher.

2.2 MFRB method

2.2.1 Bi-classification MFRB method

Multiple fitting regression (MFR) was originally proposed for spectral multiple regression analysis,[28] which established a regression model to reflect the relationship between a spectrum

and an analytic concentration. MFR can be applied to both linear multiple regression and nonlinear multiple regression.

Given the data set $D=\{\boldsymbol{x_i}, y_i\}_{i=1}^{n}$, where $\boldsymbol{x_i} \in \boldsymbol{R^n}$ is a feature vector and $y_i \in \boldsymbol{R}$, the main purpose of MFR is to use the combination of fitting functions of various kernel functions to predict y. The formula is

$$y(\boldsymbol{x}) = \sum_{i=1}^{n} a_i k_i, \tag{4}$$

$$k_i = \begin{cases} e^{-\frac{\|\boldsymbol{x}-\boldsymbol{x_i}\|^2}{c^2}} & \text{Gausian kernel,} \\ \frac{\boldsymbol{x} * \boldsymbol{x_i}^T}{c} & \text{Linear kernel,} \end{cases} \tag{5}$$

where k_i is the ith fitting function and a_i is its combination coefficient.

The combination coefficient of the Gaussian kernel is

$$\left[\boldsymbol{a_1} \ldots \boldsymbol{a_n}\right]^T = (\boldsymbol{K} + \lambda \boldsymbol{I})\boldsymbol{Y}, \tag{6}$$

$$\boldsymbol{K} = \begin{bmatrix} e^{-\frac{\|\boldsymbol{x_1}-\boldsymbol{x_1}\|^2}{c^2}} & \cdots & e^{-\frac{\|\boldsymbol{x_1}-\boldsymbol{x_n}\|^2}{c^2}} \\ \vdots & \ddots & \vdots \\ e^{-\frac{\|\boldsymbol{x_n}-\boldsymbol{x_1}\|^2}{c^2}} & \cdots & e^{-\frac{\|\boldsymbol{x_n}-\boldsymbol{x_n}\|^2}{c^2}} \end{bmatrix}. \tag{7}$$

The combination coefficient of the linear kernel is

$$\left[\boldsymbol{a_1} \ldots \boldsymbol{a_n}\right]^T = (\boldsymbol{K} + \lambda \boldsymbol{I})\boldsymbol{Y}, \tag{8}$$

$$\boldsymbol{K} = \begin{bmatrix} \frac{\boldsymbol{x_1} * \boldsymbol{x_1}^T}{c} & \cdots & \frac{\boldsymbol{x_1} * \boldsymbol{x_n}^T}{c} \\ \vdots & \ddots & \vdots \\ \frac{\boldsymbol{x_n} * \boldsymbol{x_1}^T}{c} & \cdots & \frac{\boldsymbol{x_n} * \boldsymbol{x_n}^T}{c} \end{bmatrix}, \tag{9}$$

where $\boldsymbol{I}$ refers to an $n \times n$ identity matrix, $\boldsymbol{Y} = \left[\boldsymbol{y_1}, \ldots, \boldsymbol{y_n}\right]^T$, and λ is the biased parameter. To achieve the best fitting effect of MFR, the parameters c and λ must be optimized with the given data set. Linear kernel functions are often more appropriate for large sparse matrices.

We aim to use MFR for feature extraction. For bi-classification, the labels of positive and negative classes are set to 1 and −1, respectively. The MFR is built using the features as independent variables and the class labels as dependent variables. Then, the predicted values of positive samples fluctuate around 1 and the predicted values of negative samples fluctuate around −1. This enables us to separate the samples of different classes on the basis of the predicted values. Therefore, MFR can be used as a feature extraction method that compresses original multidimensional features into a 1D integrated feature. The parameters of MFR, c and λ, are optimized by minimizing the following objective function:

$$f(c,\lambda) = \min\left(-\left|\frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2}\right|\right), \tag{10}$$

where μ_1 and σ_1 are the mean and variance of the predicted values for the positive training samples, and μ_2 and σ_2 are the mean and variance of the predicted values for the negative training samples, respectively. The parameters c and λ can be optimized using the simulated annealing algorithm or grid search technique. Equation (11) is used to extract an integrated feature using MFR that is highly similar within the same class and highly dissimilar in different classes.

To obtain a probability output for a soft decision, we assume that the predicted value of each class obeys the Gaussian distribution. The latter has a probability density function with the highest density in the center and decreasing density with increasing distance from the center, making it coincident with the distribution pattern of each class in the mapped space formed by MFR. The probability density function of the positive class is expressed as

$$p(\hat{y}|\omega_1) = \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{(\hat{y}-\mu_1)^2}{2{\sigma_1}^2}\right), \tag{11}$$

where $\hat{y}$ is the predicted value given by MFR and ω_1 indicates the positive class. Likewise, the probability density function of the negative class is expressed as

$$p(\hat{y}|\omega_{-1}) = \frac{1}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{(\hat{y}-\mu_2)^2}{2{\sigma_2}^2}\right), \tag{12}$$

where $\hat{y}$ is the predicted value given by MFR and ω_{-1} indicates the negative class. The posterior probability of an unknown sample with prediction $\hat{y}_u$ belonging to class ω_c (either class $c = 1$ or $c = -1$) is given by the Bayes formula:

$$p(\omega_c|\hat{y}_u) = \frac{p(\hat{y}_u|\omega_c)\times p(\omega_c)}{p(\hat{y}_u|\omega_1)\times p(\omega_1) + p(\hat{y}_u|\omega_{-1})\times p(\omega_{-1})}. \tag{13}$$

The prior probability $p(\omega_c)$ is calculated as follows:

$$p(\omega_1) = \frac{I_1}{I_1 + I_{-1}}, \tag{14}$$

$$p(\omega_{-1}) = \frac{I_{-1}}{I_1 + I_{-1}}, \tag{15}$$

where I_1 and I_{-1} indicate the numbers of positive and negative samples, respectively. The unknown sample is assigned to the class with the larger posterior probability as determined by the Bayesian decision rule. To avoid overfitting the training set, *K*-fold cross-validation is used, where the training set is split into *K* parts. Each of the *K* MFR models is trained on permutations of K_{-1} out of *K* parts, and $\hat{y}_i$ is evaluated on the remaining part. The union of all *K* sets of $\hat{y}_i$ forms the training set to estimate the parameters μ_i and σ_i of the Bayes model (and can also be used to adjust the MFR model parameters *c* and λ).

2.2.2 Multi-classification MFRB method

The MFRB method was first proposed to solve the problem of bi-classification. Of course, it can also be used for problems with multi-classification. In machine learning, a multi-classification problem can often be solved by decomposing it into many bi-classification problems. The decomposition scheme chosen here is a one-to-many (OVA) method. In this way, the applicability of the MFRB method can be extended to solve the multi-classification problem. In the OVA method, an *m*-meta classification problem is decomposed into *m* bi-classification problems. One class is treated as a positive class and the other classes are treated as a negative class. Each binary classification problem is solved by the MFRB bi-classification method, which gives the probability of each test set sample belonging to each label, so the result is presented in the form of *m* posterior probabilities:

$$\sum_{i=1}^{m} p\left(\omega_i \middle| \hat{y}(\boldsymbol{x})\right) = 1, \tag{16}$$

where $p\left(\omega_i \middle| \hat{y}(\boldsymbol{x})\right)$ is the posterior probability of the sample of the unknown tag in the class *I* positive class MFRB classification. Thus, we normalize the probability

$$p\left(\omega_i \middle| \boldsymbol{x}\right) = \frac{p\left(\omega_i \middle| \hat{y}_i(\boldsymbol{x})\right)}{\sum_{j=1}^{m} p\left(\omega_j \middle| \hat{y}_j(\boldsymbol{x})\right)}. \tag{17}$$

The unknown sample *X* is assigned to the category with the higher posterior probability:

$$\text{class} = \arg\max p\left(\omega_i \middle| \boldsymbol{x}\right) \quad i = 1, \ldots, m. \tag{18}$$

3. Comparison of Classification Effects of Two Methods on Different Types of Single-cell RNA Data

In this paper, two different types of single-cell RNA data are used for experiments, namely, large-scale single-cell RNA data sets and unbalanced small-sample single-cell RNA data sets. The advantages and disadvantages of the MFRB and PSVM methods in single-cell RNA data classification are analyzed. The PSVM algorithm is obtained from the LIBLINEAR toolkit and programmed by referring to the MFRB method. All experimental codes are implemented through MATLAB2016.

3.1 Classification of large-scale data sets

3.1.1 Experimental purpose

When the number of experimental samples is large, the sample data of various training sets will be relatively rich; thus, the training of the model will be more sufficient and the classification algorithm can achieve relatively good performance. For single-cell RNA data, somatic cells of some tissues are easy to obtain and detect, so data sets with relatively large sample sizes can be obtained. Therefore, in Experiment 1 (Exp 1), the data of human embryonic stem cells are selected from a large-scale single-cell RNA data set to test the recognition performance of the two algorithms.

3.1.2 Experimental design

Exp 1 uses human embryonic stem cell GSE64016 data,[29] which consist of sequencing data of 247 H1-Fucci single cells. The data are divided into three categories: G1, G2-M, and S. Among them, the G2-M category has 76 samples (category label 1), the S category has 80 samples (category label 2), and the G1 category has 91 samples (category label 3). Among the 247 samples, the data contain 19084 genes, i.e., the dimension is 19084. During the experiment, 70% of the samples from the data set are randomly selected from each category to form the test set, and the remaining 30% form the verification set. The complete experiment, including the process of randomly dividing the data set, is repeated 30 times. The average values of the three evaluation indexes of accuracy (ACC), sensitivity (SN), and specificity (SP) of the 30 classification results are taken as the final result of the experiment. When calculating SN and SP of the first category, the first category is taken as the positive category, and the other categories are taken as the negative category. When calculating SN and SP of the second category, the second category is taken as the positive category, and the other categories are taken as the negative category. The same applies to the calculation of other categories of SN and SP. By looking at SN and SP of each category, we can assess the accuracy of the classifier for each category, or which categories the classifier tends to misidentify as the same category.

A linear kernel function is used in the PSVM method. The MFRB method uses Gaussian and linear kernel functions to generate the contrast between the two types of kernel functions. The

MFRB method is used to find the optimal parameters (c, λ), and the lattice search space is $\{10^{-10},10^{-9},\ldots,10^{9},10^{10}\}\times\{10^{-10},10^{-9},\ldots,10^{9},10^{10}\}$. For the MFRB method, the process of finding the optimal value uses fivefold cross-validation for each pair of (c, λ) parameters. For the PSVM method, the optimal parameter c is found in the space of $\{-5,-4,-3,\ldots,3,4,5\}$, and the rest of the parameters are the default parameters of the LIBLINEAR toolbox.

3.1.3 Results and discussion

First, the distribution characteristics of three types of data are studied for the data sets. The T-SNE method, a nonlinear dimension reduction method commonly used in single-cell RNA data processing (i.e., t-distribution adjacent embedding method), is used. The programming uses the TSNE function in the DRToolbox toolkit of MATLAB. The results are shown in Fig. 1.

Figure 1 shows the profile of single-cell RNA data of human embryonic stem cells after dimension reduction. The red dots are G2-M cells, the blue dots are S cells, and the green dots are G1 cells. As can be seen from Fig. 1, the three categories are relatively distinct in space. The red dots refer to the category labeled 1, and the samples of the category labeled 3 are relatively dispersed in space. The green dots refer to the category labeled 3, and the samples of the category labeled 3 are relatively compact in space. There is an intersection between categories labeled 1 and 2.

According to the experimental results given in Table 1, the overall classification accuracies of the PSVM, MFRB (Gaussian nucleus), and MFRB (linear nucleus) methods in Exp 1 were all above 90%, indicating that they can achieve very good results in the classification of human embryonic stem cell data. The linear kernel MFRB method has two percentage points higher classification accuracy than the other two methods, indicating better classification performance. For label 1, it can be seen from the SN value that the classification accuracy of the PSVM

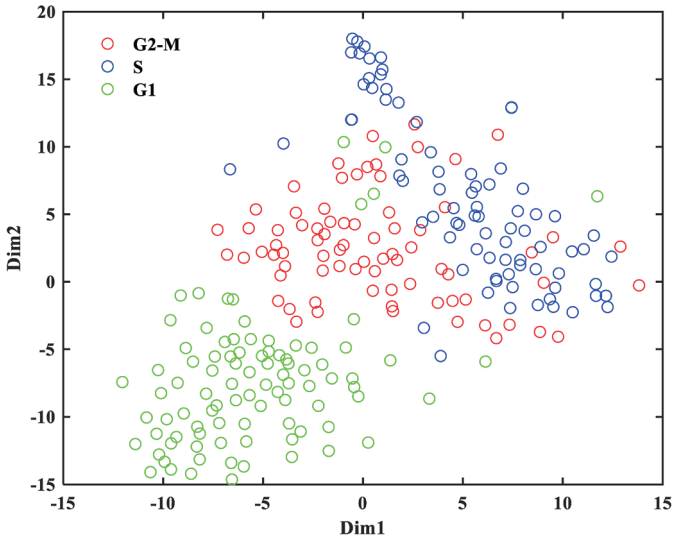

Fig. 1. (Color online) Dimension-reduced profile of single-cell RNA data of human embryonic stem cells.

Table 1
Results of classification of single-cell RNA data of human embryonic stem cells by three algorithms.

Category labels	ACC			SN			SP		
	PSVM	MFRB Gaussian kernel	MFRB Linear kernel	PSVM	MFRB Gaussian kernel	MFRB Linear kernel	PSVM	MFRB Gaussian kernel	MFRB Linear kernel
1	0.92	0.92	0.94	0.94	0.91	0.91	0.93	0.95	0.96
2				0.91	0.89	0.95	0.96	0.96	0.96
3				0.92	0.95	0.95	0.99	0.96	0.99

method of 94% is higher than that of 91% for the MFRB method. For the classification of label 2, the SN values show that the PSVM method and the Gaussian kernel MFRB method have similar classification performances, but the linear kernel MFRB method has superior performance to both of them. For the classification of label 3, it can be seen from the SN value that the classification performance of the MFRB method is better than that of the PSVM method, with the SN value increased by three percentage points. The linear kernel MFRB method and the Gaussian kernel MFRB method show little difference for sample data labels 1 and 3, which is related to the characteristics of the data itself. The classification results are consistent with the data distribution.

3.2 Classification of small-sample unbalanced data sets

3.2.1 Experimental purpose

Under the condition of small-sample data, the number of sample data for some categories is very small, and an extreme data imbalance may occur. This situation often leads to the overfitting of one or more of the data sets in the trained model. Although the accuracy of such training is not low, it may not be able to distinguish all types of data. In the single-cell RNA data, the number of some cells is very small, so the sample data of the training set will be relatively scarce, and the data set will be unbalanced when classified relative to other cells. Relatively speaking, it is difficult to obtain sufficient information from a limited number of samples to make the classification algorithm achieve relatively good results. Thus, modeling such single-cell RNA data is a major challenge. For example, some embryonic stem cells and cancer cells are relatively few, and some abnormal cells are rare. Therefore, it is of practical significance to study small sample sizes and unbalanced data sets. In Exp 2, on the basis of the characteristics of unbalanced data sets with small samples, a group of single-cell RNA data of epithelial cells is selected for classification.

3.2.2 Experimental design

Exp 2 uses single-cell RNA data of epithelial cells.[30] Table 2 shows the information of the epithelial single-cell RNA data. In each experiment, 70% of the samples of the data set are

Table 2
Information of epithelial single-cell RNA data.

Category labels	1	2	3	4	5	6	7	8	9	10
Sample size	48	14	12	14	15	24	9	7	7	10
Feature dimension	55182									

randomly selected from each category to form the test set, and the remaining 30% are selected to form the verification set.

The evaluation indexes of the experimental results are the same as those of Exp 1. Firstly, the experiment is carried out under the normal experimental classification. Considering that a lack of samples may lead to unsatisfactory results, the single-cell RNA data filling method is used to improve the data status and enhance the classification performance. Therefore, in this study, we use the data after migration to carry out the classification experiment. The data filling algorithm here adopts a random walking algorithm[31] that connects genes with a similar topological structure through network propagation and updates the data through iterative optimization to overcome the noise problem caused by the "drop out" phenomenon of the single-cell RNA data. Other experimental parameters, steps, and so forth are kept unchanged. The two experiments are repeated 30 times before and after data filling. The average values of ACC, SN, and SP of 30 classification results are taken as the final results of the experiment.

Moreover, considering that the difficulty of classification may be aggravated by the small number of samples and a large number of categories in the data set, in this study, we remove the category of data with the least number of samples each time based on the original data set, so that it does not participate in the classification experiment, and the remaining experimental parameters and steps are kept unchanged. In accordance with the data sets of different categories, including the process of randomly dividing the training set and test set, each test is repeated 30 times. The average values of ACC, SN, and SP of 30 classification results are taken as the final results of this test.

3.2.3 Results and discussion

Firstly, the distribution characteristics of the ten types of data are studied for each data set. The T-SNE method, which is common in single-cell RNA data processing, is used. The results are shown in Figs. 2 and 3.

Figure 2 shows the 2D spatial distribution of single-cell RNA data of epithelial cells after dimension reduction. Figure 3 shows the 2D spatial distribution of single-cell RNA data of epithelial cells processed using the random walking algorithm. Red, green, blue, turquoise, purple, yellow, and black dots and green, blue-green, and brown crosses represent epithelial cells labeled 1 to 10, respectively.

As can be seen from Figs. 2 and 3, most of the samples of the category labeled 1 are distributed sporadically, and the number of samples is small. Moreover, each category is widely distributed in space with almost no obvious aggregation or dividing line. Other dimension reduction methods also give similar distribution results. It can be seen that there is no clear

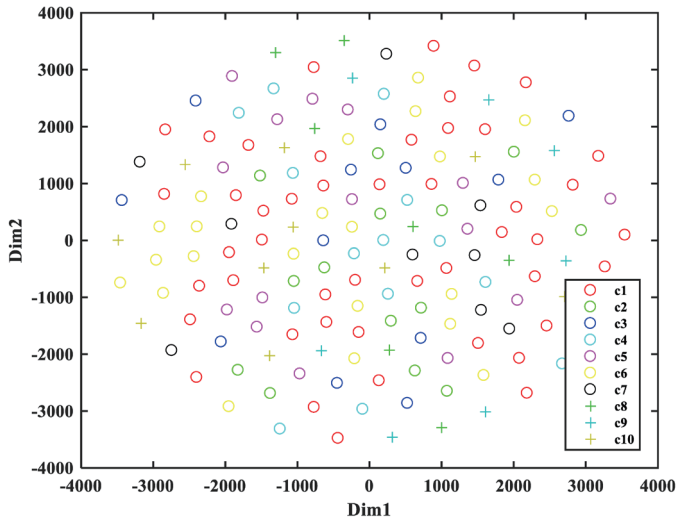

Fig. 2. (Color online) Dimension-reduced profile of single-cell RNA data of epithelial cells.

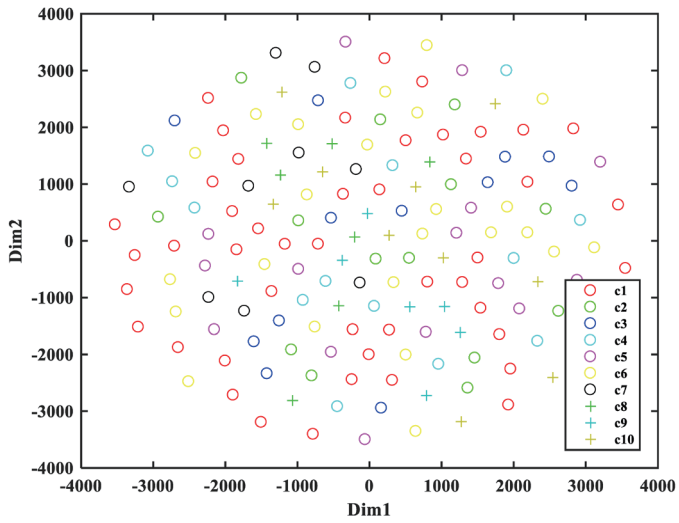

Fig. 3. (Color online) Dimension-reduced profile of epithelial cells after random walking.

distinction between the categories of data in this data set in the case of a low number of dimensions.

As can be seen from the accuracy values in Table 3, the accuracy of the PSVM method is about 50%, whereas that of the MFRB method is about 30%. Compared with the MFRB method, the PSVM method has a better classification performance. The accuracy of the MFRB method is about 35% with the Gaussian kernel and 22% with the linear kernel.

When the random walking algorithm was not used, the observation of the SN value showed that the sensitivity of the PSVM method to the data of Tags 1, 4, 7, and 8 was lower than 55%. In particular, the specificity of Tag 1 was only about 23%, indicating that the PSVM method could not correctly classify these tags. The Gaussian kernel MFRB method has a sensitivity of more

Table 3
Results of classification of single-cell RNA data of epithelial cells by three algorithms.

Category labels	ACC			SN			SP		
	PSVM	MFRB Gaussian kernel	MFRB Linear kernel	PSVM	MFRB Gaussian kernel	MFRB Linear kernel	PSVM	MFRB Gaussian kernel	MFRB Linear kernel
1	0.55	0.35	0.21	0.24	0.34	0.00	1.00	0.96	1.00
2				0.90	0.43	0.43	0.85	0.96	0.93
3				0.78	0.13	0.37	0.94	0.89	0.88
4				0.48	0.17	0.30	0.96	0.87	0.95
5				0.84	0.70	0.38	1.00	0.92	0.86
6				0.72	0.66	0.17	0.96	0.73	0.98
7				0.43	0.30	0.15	0.99	0.94	0.98
8				0.45	0.00	0.35	1.00	1.00	0.85
9				0.73	0.15	0.30	0.95	0.99	0.87
10				0.64	0.00	0.37	0.86	1.00	0.85

than 60% to only the data of Tags 5 and 6, and the overall classification performance is poor. The linear kernel MFRB method has no more than 50% sensitivity to each category that can be correctly classified.

As can be seen from Table 4, after the random walking algorithm was used to process the data, the total accuracy of the PSVM method decreased by about five percentage points, whereas that of the MFRB method remained unchanged. This shows that the random walking algorithm does not affect the overall accuracy of data. The sensitivities of the PSVM and MFRB methods to the data after the random walking were significantly different from those before the random walking. By observing the sensitivity data of each category of the MFRB algorithm based on the Gaussian kernel after the random walking, it can be seen that filling such small-sample data with the random walking algorithm cannot significantly improve the distribution characteristics of all types of data, nor can it improve the classification results for unbalanced small-sample data.

As can be seen from Fig. 4, the accuracies of the PSVM method and Gaussian kernel MFRB method increase with decreasing number of data sets: the accuracy of the PSVM method increases from 0.5 to 0.55 and that of the Gaussian kernel MFRB method increases from 0.3 to 0.38. This suggests that reducing the number of categories can improve the classification results when the data sets are unbalanced and the sample size is small. The accuracy of the linear kernel MFRB method is almost unchanged, indicating that it is not affected by the number of categories.

This experiment is repeated five times, with the single-cell RNA data of epithelial cells reduced stepwise from ten to six categories (with the category having the smallest sample size removed each time). The sensitivity of the PSVM method changes with the number of categories. Because the sample size is small, the sensitivity fluctuates, but the overall sensitivity of each category increases with decreasing number of categories. The classification of each category was improved by reducing the number of categories. From the observed change in the sensitivity of the MFRB method, it can be seen that the sensitivity of each category is affected differently by the decrease in the number of categories, with the sensitivity of some categories increasing significantly and that of others decreasing significantly. Even if reducing the number of

Table 4
Results of classification of single-cell RNA data of epithelial cells after random walking by three algorithms.

Category labels	ACC			SN			SP		
	PSVM	MFRB Gaussian kernel	MFRB Linear kernel	PSVM	MFRB Gaussian kernel	MFRB Linear kernel	PSVM	MFRB Gaussian kernel	MFRB Linear kernel
1	0.50	0.34	0.24	0.13	0.67	0.00	1.00	0.77	1.00
2				0.83	0.55	0.63	0.83	0.81	0.78
3				0.67	0.20	0.47	0.94	0.97	0.90
4				0.55	0.23	0.55	0.95	0.78	0.75
5				0.86	0.28	0.45	1.00	0.97	0.96
6				0.64	0.17	0.09	0.95	0.98	0.97
7				0.50	0.00	0.10	0.99	0.96	0.98
8				0.40	0.00	0.35	0.99	0.99	1.00
9				0.68	0.00	0.25	0.89	1.00	0.97
10				0.68	0.00	0.33	0.92	1.00	0.87

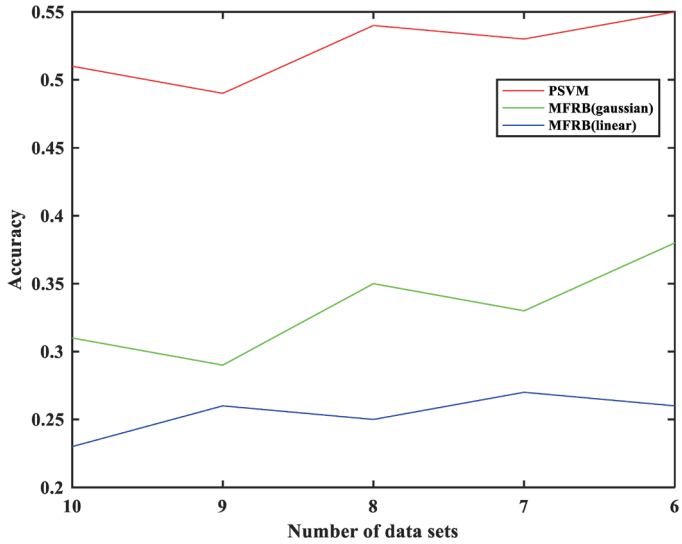

Fig. 4. (Color online) Relationship between data set size and accuracy rate.

categories can improve the problem of unbalanced data, the Gaussian kernel MFRB method will still be affected by unbalanced data. The sensitivity of the linear kernel MFRB method only fluctuates around the mean value and is not affected by the change in the number of categories. The effects of the two methods for each tag were described in the previous sections and will not be repeated here.

With the continuous advancement of single-cell sequencing technology, supplementing and perfecting single-cell data mining algorithms will help to understand the heterogeneity of single cells and have great significance and value for the precise prevention, diagnosis, and treatment of diseases. Machine learning and data mining are currently the preferred technologies for big data information processing. With the assistance of AI algorithms to improve the accuracy of single-cell sequencing data analysis, in this study, we adopt the widely used SVM method in

machine learning and the newly developed MFRB methods[25] for comparative research. We take three types of single-cell data, namely, large-scale single-cell data sets, small-sample unbalanced single-cell data sets, and single-cell data sets with batch effects, as examples.[26] Therefore, in three situations, we analyze the performance of the two classification algorithms and gain an in-depth understanding of the distribution characteristics of single-cell data. The MFRB method is a new algorithm that combines the MFR method with Bayesian decision-making in machine learning. The PSVM method performs well in data classification.

4. Conclusion

In this paper, for large-scale data sets and small unbalanced data sets, the PSVM method and the Gaussian kernel and linear kernel MFRB methods were used to conduct experiments to analyze their performance in the classification of single-cell RNA sequencing data. For the classification of large-scale data sets, all three methods have high accuracy. For the unbalanced data set used in the experiment, the addition of the random walking algorithm has a slightly negative impact on the experimental results. Before adding this algorithm, the PSVM method and the Gaussian kernel and linear kernel MFRB methods have different sensitivities to the data under each label, but these sensitivities are greater than 0. After the addition of the random walking algorithm, the sensitivity of the Gaussian kernel MFRB method drops to 0 for several categories, indicating that the addition of the random walking algorithm increases the effect of the unbalanced data set on the performance of the Gaussian kernel MFRB method. Reducing the number of categories can improve the accuracy of the PSVM and Gaussian kernel MFRB methods. The PSVM method increases the sensitivity of each tag as the number of categories decreases. With the decrease in the number of categories, the sensitivity of each label decreases, and the overall accuracy rate increases gradually. The classification results of the linear kernel MFRB method are not affected by the number of classes. Although the SVM method, which is the most commonly used method, and the newly developed MFRB methods with excellent performance are adopted in this study, the classification performance cannot be guaranteed to be the best among all existing classification algorithms. We hope to explore more types of single-cell RNA sequencing data and classification algorithms to compare results. For unbalanced small-sample data, the main reason why it is difficult to achieve a high recognition rate with a small number of samples is still unclear, and the corresponding solution is also still unclear. These problems will be gradually addressed in future work.

Acknowledgments

This work was supported by Longyan University's Qi Mai Science and Technology Innovation Fund Project of Shanghang County (2017SHQM07) and the Great Project of Production, Teaching, and Research of Fujian Provincial Science and Technology Department (2019H6023).

References

1 Y. Lin, P. Yang, Y. Chen, J. Zhu, X. Zhang, and C. Ma: Arch Gynecol Obstet. **299** (2019) 559. https://doi.org/10.1007/s00404-018-4958-3
2 N. Navin, J. Kendall, and J. Troge: Nature **472** (2011) 90. https://doi.org/10.1038/nature09807
3 M. Sarimollaoglu, D. A. Nedosekin, E. I. Galanzha, and V. P. Zharov: Proc. SPIE 8581, Photons Plus Ultrasound: Imaging and Sensing (SPIE, 2013) 858124. https://doi.org/10.1117/12.2007964
4 L. Huang and R. T. Kennedy: Trends Anal. Chem. **14** (1995) 158. https://doi.org/10.1016/0165-9936(95)98313-W
5 M. Shi, Z. H. Song, X. H. Geng, D. P. Wu, and Y. F. Guan: Chin. J. Chromatogr. **35** (2017) 105. https://doi.org/10.3724/SP.J.1123.2016.08039
6 Dalian Institute of Chemical Physics: Research Progress in Quantitative Analysis of High Throughput Multiple Proteomics [R] (Dalian: China Academy of Sciences, 2013). http://www.dicp.ac.cn
7 S. C. Bendall, E. F. Simonds, P. Qiu, El-Ad D. Amir, P. O. Krutzik, R. Finck, R. V. Bruggner, R. Melamed, and A. Trejo: Science **332** (2011) 687. https://doi.org/10.1126/science.1198704
8 J. Liu and Z. Liu: Chin. J. Chromatogr. **34** (2016) 1154. https://doi.org/10.3724/SP.J.1123.2016.08041
9 G. L. Peng, J. H. Chen, and D. J. Rong: Rural Economy Sci. Technol. **10** (2017) 120.
10 Y. Liu and B. Q. Wu: Chin. J. Pathol. **40** (2011) 718. https://doi.org/10.3760/cma.j.issn.0529-5807.2011.10.022
11 W. J. Wang: Nat. Med. J. Chin. **95** (2015).
12 C. Zong, S. Lu, A. R. Chapman, and X. S. Xie: Science **338** (2012) 1622. https://doi.org/10.1126/science.1229164
13 M. D. Williams, R. Reeves, L.S. Resar, and H. H. Hill Jr.: Anal. Bioanal. Chem. **405** (2013) 5013. https://doi.org/10.1007/s00216-013-6777-5
14 Z. X. Zhu and X. Chen: Genomics Appl. Biol. **5** (2016) 000902. https://www.cnki.net/kcms/doi/10.13417/j.gab.034.000902.html
15 J. K. Cheng, W. H. Huang, and Z. L. Wang: Chin. J. Chromatogr. **25** (2007) 1.
16 Y. Han: Sci. Chin. **20** (2016) 43.
17 H. Chitsaz, J. L. Yee-Greenbaum, G. Tesler, M.J. Lombardo, C. L Dupont, J. H. Badger, M. Novotny, D. B. Rusch, L. J. Fraser, N. A. Gormley, O. Schulz-Trieglaff, G. P. Smith, D. J. Evers, P. A. Pevzner, and R. S. Lasken: Nat. Biotechnol. **10** (2011) 915. https://www.nature.com/articles/nbt.1966
18 S. Miyatake, E. Koshimizu, A. Fujita, R. Fukai, E. Imagawa, C. Ohba, I. Kuki, M. Nukui, A. Araki, Y. Makita, T. Ogata, M. Nakashima, Y. Tsurusaki, N. Miyake, H. Saitsu, and N. Matsumoto: J. Human Genet. **60** (2015) 175. https://doi.org/10.1038/jhg.2014.124
19 S. Vinga and J. S. Almeida: J. Theor. Biol. **231** (2004) 377. https://doi.org/10.1016/j.jtbi.2004.06.030
20 D. Koslicki: Bioinformatics (Oxford, England) **27** (2011) 1061. https://doi.org/10.1093/bioinformatics/btr077
21 L. W. Plasschaert, R. Žilionis, and R. Choo-Wing: Nature **560** (2018) 377. https://www.nature.com/articles/s41586-018-0394-6
22 D. T. Montoro, A. L. Haber, M. Biton, V. Vinarsky, B. Lin, S. E. Birket, F. Yuan, S. Chen, H. M. Leung, J. Villoria, N. Rogel, G. Burgin, A. M. Tsankov, A. Waghray, M. Slyper, J. Waldman, L. Nguyen, D. Dionne, O. Rozenblatt-Rosen, P. R. Tata, H. Mou, M. Shivaraju, H. Bihler, M. Mense, G. J. Tearney, S. M. Rowe, J. F. Engelhardt, A. Regev, and J. Rajagopal: Nature **560** (2018) 319. https://doi.org/10.1038/s41586-018-0393-7
23 P. A. Northcott, I. Buchhalter, A. S. Morrissy, M.-L. Yaspo, R. Kriwacki, A. Gajjar, J. H. Zhang, R. Beroukhim, E. Fraenkel, J. O. Korbel, B. Brors, M. Schlesner, R. Eils, M. A. Marra, S. M. Pfister, M. D. Taylor, and P. Lichter: Nature **547** (2017) 311. https://doi.org/10.1038/nature22973
24 Y. L. Hu, T. Hase, H. P. Li, S. Prabhakar, H.i Kitano, S. K. Ng, S. Ghosh, and L. J. K. Wee: BMC Genomics **17** (2016) 1025. https://doi.org/10.1186/s12864-016-3317-7
25 G. Z. Huang, M. S. Yuan, M. L. Chen, S. Prabhakar, H. Kitano, S. K. Ng, S. Ghosh, and L. J. K. Wee: Analyst **142** (2017) 3588. https://doi.org/10.1039/C7AN00944E
26 C. W. Hsu, C. C. Chang, and C. J. Lin: A Practical Guide to Support Vector Classification [EB/OL] (2003). http://www.csie.ntu.edu.tw/~cjlin
27 J. C. Platt: Adv. Large Margin Classifiers **10** (1999) 61. https://www.researchgate.net/publication/2594015
28 X. J. Chen, Y. J. Lai, X. Chen, Y. J. Shi, and D. H. Zhu. Analyst **141** (2016) 5759. https://doi.org/10.1039/c6an01201a
29 N. Leng, L. F. Chu, C. Barry, Y. Li, J. Choi, X. M. Li, P. Jiang, R. M. Stewart, J. A. Thomson, and C. Kendziorski: Nat. Methods **12** (2015) 947. https://doi.org/10.1038/nmeth.3549
30 H. P. Li, E. T. Courtois, D. Sengupta, Y. Tan, K. H. Chen, J. J. L. Goh, S. L. Kong, C. Chua, L. K. Hon, W. S. Tan, M. Wong, P. J. Choi, L. J. K. Wee, A. M. Hillmer, I. B. Tan, P. Robson, and S. Prabhakar: Nat. Genetics **49** (2017) 708. https://doi.org/10.1038/ng.3818
31 L. Cowen, T. Ideker, B. J. Raphael, and R. Sharan: Nat. Rev. Genet. **18** (2017) 551. https://doi.org/10.1038/nrg.2017.38