Selecting Classification Methods for Small Samples of Next-Generation Sequencing Data

Jiadi Zhu[1], *Ziyang Yuan*[2], *Lianjie Shu*[3], *Wenhui Liao*[4]*, *Mingtao Zhao*[5]* *and Yan Zhou*[2]*

[1] Department of Mathematics and Statistics, Xidian University, Xi'an, China, [2] Shenzhen Key Laboratory of Advanced Machine Learning and Applications, College of Mathematics and Statistics, Institute of Statistical Sciences, Shenzhen University, Shenzhen, China, [3] Faculty of Business Administration, University of Macau, Macau, China, [4] GuangDong University of Finance, Guangzhou, China, [5] Institute of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu, China

OPEN ACCESS

Edited by:
Zhonghua Liu,
The University of Hong Kong,
Hong Kong

Reviewed by:
Qi Shen,
Icahn School of Medicine at Mount Sinai, United States
Xiaobo Guo,
Sun Yat-sen University, China

***Correspondence:**
Wenhui Liao
lwh_gduf@sina.com
Mingtao Zhao
mingtao.zhao@outlook.com
Yan Zhou
zhouy1016@szu.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *15 December 2020*
Accepted: *01 February 2021*
Published: *04 March 2021*

Citation:
Zhu J, Yuan Z, Shu L, Liao W, Zhao M and Zhou Y (2021) Selecting Classification Methods for Small Samples of Next-Generation Sequencing Data. Front. Genet. 12:642227. doi: 10.3389/fgene.2021.642227

Next-generation sequencing has emerged as an essential technology for the quantitative analysis of gene expression. In medical research, RNA sequencing (RNA-seq) data are commonly used to identify which type of disease a patient has. Because of the discrete nature of RNA-seq data, the existing statistical methods that have been developed for microarray data cannot be directly applied to RNA-seq data. Existing statistical methods usually model RNA-seq data by a discrete distribution, such as the Poisson, the negative binomial, or the mixture distribution with a point mass at zero and a Poisson distribution to further allow for data with an excess of zeros. Consequently, analytic tools corresponding to the above three discrete distributions have been developed: Poisson linear discriminant analysis (PLDA), negative binomial linear discriminant analysis (NBLDA), and zero-inflated Poisson logistic discriminant analysis (ZIPLDA). However, it is unclear what the real distributions would be for these classifications when applied to a new and real dataset. Considering that count datasets are frequently characterized by excess zeros and overdispersion, this paper extends the existing distribution to a mixture distribution with a point mass at zero and a negative binomial distribution and proposes a zero-inflated negative binomial logistic discriminant analysis (ZINBLDA) for classification. More importantly, we compare the above four classification methods from the perspective of model parameters, as an understanding of parameters is necessary for selecting the optimal method for RNA-seq data. Furthermore, we determine that the above four methods could transform into each other in some cases. Using simulation studies, we compare and evaluate the performance of these classification methods in a wide range of settings, and we also present a decision tree model created to help us select the optimal classifier for a new RNA-seq dataset. The results of the two real datasets coincide with the theory and simulation analysis results. The methods used in this work are implemented in the open-scource R scripts, with a source code freely available at https://github.com/FocusPaka/ZINBLDA.

Keywords: RNA-seq data, classification, PLDA, NBLDA, ZIPLDA, ZINBLDA

1. INTRODUCTION

RNA sequencing (RNA-seq), which involves directly sequencing complementary DNAs and aligning the sequences to the reference genome or transcriptome, has emerged as a powerful technology for measuring gene expression (Mardis, 2008; Morozova et al., 2009; Wang et al., 2009). In recent years, the affordability and effectiveness of RNA-seq has resulted in its application in biological and medical studies, such as genomics research (Nagalakshmi et al., 2008; Trapnell et al., 2010) and clinical use (Berger et al., 2010; Biesecker et al., 2012). Unlike microarray technology, RNA-seq allows for the detection of novel transcripts with low background signals. One of the biological applications of RNA-seq is inferring differential expression (DE) genes between different conditions or tissues. Existing popular methods include edgeR (Robinson and Smyth, 2008; Robinson et al., 2010), DESeq2 (Love et al., 2014), and LFCseq (Lin et al., 2014). Another important application is the diagnosis of diseases. Numerous discriminant methods have been proposed for the diagnosis of diseases using microarray data, such as diagonal linear discriminant analysis and diagonal quadratic discriminant analysis in Dudoit et al. (2002). In previous RNA-seq experiments, the read counts (the number of short reads mapped to the reference genome) have been used to measure the expression level. However, because the expression matrix entries are non-negative integers, classification methods that follow a Gaussian distribution may not perform well for RNA-seq data.

Classification methods based on different discrete distributions have been proposed for RNA-seq data. Witten (2011) assumed RNA-seq data follow a Poisson distribution and proposed a Poisson linear discriminant analysis (PLDA) method. Comparison studies (Tan et al., 2014) have shown that PLDA performs much better than the method used for microarray data when classifying RNA-seq data. Considering the overdispersion of RNA-seq data, Dong et al. (2016) assumed that data follow a negative binomial distribution and developed a negative binomial linear discriminant analysis (NBLDA) method. Zhou et al. (2018) found excess zeros in real RNA-seq data and proposed a zero-inflated Poisson logistic discriminant analysis (ZIPLDA) method, which assumes RNA-seq data follow a mixture distribution with a point mass at zero and a Poisson distribution.

Due to the shallow sequence depth and dispersed biological replicates, there may be excess zeros and overdispersion in a real RNA-seq dataset, which should be considered when conducting data analysis. For instance, the real dataset TCGA-LIHC, which includes a cancerous and normal group, contains about 43.24% zeros of all numerical values, and the estimated dispersion parameter is 1.12. Therefore, a natural assumption would be to extend the existing discrete distribution to a mixture distribution with a point mass at zero and a negative binomial distribution. We call this method zero-inflated negative binomial logistic discriminant analysis (ZINBLDA). To obtain the model, which is similar to ZIPLDA, we built a mixture distribution with a point mass at zero and a negative binomial distribution for the remaining data. We then estimated the parameters in the model. Finally, we obtained a classifier by Bayes rule to predict for a future observation. We also analyzed the relationship between the above four classification methods, and the resulting discriminant scores for the four classification methods showed that they can transform into each other in some cases. We examined these four methods from the perspective of their parameters and determined how the parameters provide the link between the selected optimal method and the model classification performance. In addition, we built a decision tree to help us select the optimal classifier from these four methods for a new dataset.

The remainder of the article is organized as follows. In section 2, we review the existing three classification methods and propose the ZINBLDA method for overdispersion RNA-seq data with an excess of zeros. We also give the estimation of the parameters in the model in detail. We further discuss the transformation relations between the four methods. Section 3 discusses the results of the simulation studies that were conducted to evaluate the performance of the four methods in a wide range of settings. This section also presents a decision tree that was built to select the optimal classifier from these four methods for a new dataset. In section 4, we employ the four methods to analyze two real RNA-seq datasets and evaluate their performance. Finally, we conclude the work with a discussion of the findings and future directions.

2. CLASSIFICATION METHODS

There are three existing classification methods for RNA-seq data: PLDA (Witten, 2011), NBLDA (Dong et al., 2016), and ZIPLDA (Zhou et al., 2018). We propose a new discriminant analysis method to model overdispersion RNA-seq data with excess zeros. We examined these four methods from the perspective of their parameters and analyzed the transformation relations between the methods.

Before introducing the methods, we must first specify some notations used in this work. In this paper, K is the number of classes, and X_{ki_kg} denotes the number of read counts that are mapped to gene g in sample i_k of class k, where $k = 1, \cdots, K$; $i_k = 1, \cdots, n_k$; and $g = 1, \cdots, G$. Specifically, there are n_k samples in class k, and $n = \sum_{k=1}^{K} n_k$ denotes the total number of samples for all classes.

2.1. Principle of the Classifiers

The principle of the classifiers is applicable for the following four classifiers. Suppose that for the training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ we wished to classify a new observation $\boldsymbol{x}^* = (X_1^*, \cdots, X_G^*)^T$. If y^* is the unknown label of $\boldsymbol{x}^*$, by Bayes' rule

$$P(y^* = k|\boldsymbol{x}^*) \propto f_k(\boldsymbol{x}^*)\pi_k, \tag{1}$$

where f_k is the probability density function of an observation in class k, and π_k is the prior probability that an observation belongs to class k. In general, we can use $\pi_k = n_k/n$ to satisfy $\sum_{k=1}^{K} \pi_k$=1. We define a discriminant score function as $d_k(\boldsymbol{x}^*) = \log[P(\boldsymbol{x}^*|y^* = k)\pi_k]$ on the basis of formula (1) and assign a new observation to the class for which the discriminant score is the highest.

2.2. Poisson Linear Discriminant Analysis

For PLDA, Witten (2011) assumed that RNA-seq data follow a Poisson distribution, that is,

$$X_{ki_kg}|y_{i_k} = k \sim Poisson(\mu_{ki_kg}), \quad \mu_{ki_kg} = d_{kg}s_{i_k}\lambda_g, \tag{2}$$

where μ_{ki_kg} is the expectation for gene g in sample i_k of class k, s_{i_k} is the size factor used to identify individuals in the kth class, λ_g is the total number of read counts for gene g, and d_{kg} allows for the differential expression of gene g between the different classes. Following the expression of (2), the probability density function is

$$P(X_{ki_kg} = x_{ki_kg}) = \frac{\mu_{ki_kg}^{x_{ki_kg}}}{(x_{ki_kg})!}e^{-\mu_{ki_kg}}.$$

Thus, according to formula (1), the discriminant score of PLDA is obtained by

$$d_k(\boldsymbol{x}^*) = \sum_{g=1}^{G} X_g^* \log(d_{kg}) - s^* \sum_{g=1}^{G} \lambda_g d_{kg} + \log \pi_k + C, \tag{3}$$

where s^* is the size factor of test observation, and C represents a constant that is unrelated to the class label.

2.3. Negative Binomial Linear Discriminant Analysis

Modeling RNA-seq data with a negative binomial distribution instead of a Poisson distribution is a natural extension. Dong et al. (2016) proposed NBLDA to allow for cases where variance is greater than or equal to the mean, and they also demonstrated that NBLDA is more suitable when biological replicates are available. The negative binomial distribution is expressed as

$$X_{ki_kg}|y_{i_k} = k \sim NB(\mu_{ki_kg}, \phi_g), \quad \mu_{ki_kg} = d_{kg}s_{i_k}\lambda_g, \tag{4}$$

where ϕ_g is a non-negative dispersion parameter, and the rest of parameters are the same as for PLDA. Therefore, the probability density function of $X_{ki_kg} = x_{ki_kg}$ in model (4) is

$$P(X_{ki_kg} = x_{ki_kg}) = \frac{\Gamma(x_{ki_kg} + \phi_g^{-1})}{(x_{ki_kg})!\Gamma(\phi_g^{-1})}(\frac{\mu_{ki_kg}\phi_g}{1 + \mu_{ki_kg}\phi_g})^{x_{ki_kg}}$$
$$(\frac{1}{1 + \mu_{ki_kg}\phi_g})^{\phi_g^{-1}}.$$

Similarly, the discriminant score can be obtained by

$$\begin{aligned} d_k(\boldsymbol{x}^*) &= \sum_{g=1}^{G} X_g^*[\log(d_{kg}) - \log(1 + s^*\lambda_g d_{kg}\phi_g)] \\ &- \sum_{g=1}^{G} \phi_g^{-1} \log(1 + s^*\lambda_g d_{kg}\phi_g) + \log \pi_k + C. \end{aligned} \tag{5}$$

2.4. Zero-Inflated Poisson Logistic Discriminant Analysis

Considering data with excess zeros due to missing records or no observation signal, Zhou et al. (2018) proposed ZIPLDA method, which assumes that data follow a zero-inflated Poisson distribution. The distribution is expressed as

$$X_{ki_kg} \sim \begin{cases} \delta_{\{0\}}, & p_{ki_kg}, \\ \text{Poisson}(\mu_{ki_kg}), & (1 - p_{ki_kg}), \end{cases}$$

where $\delta_{\{0\}}$ denotes the point mass at zero, p_{ki_kg} is the probability of $\delta_{\{0\}}$ in gene g of sample i_k in class k, and μ_{ki_kg} is same as in the former two classifiers. Thus, the probability of X_{ki_kg} is written as

$$P(X_{ki_kg}) = \begin{cases} p_{ki_kg} + (1 - p_{ki_kg})e^{-\mu_{ki_kg}}, & X_{ki_kg} = 0, \\ (1 - p_{ki_kg})\frac{\mu_{ki_kg}^{X_{ki_kg}}}{(X_{ki_kg})!}e^{-\mu_{ki_kg}}, & X_{ki_kg} > 0. \end{cases}$$

Additionally, the probability density function of $X_{ki_kg} = x_{ki_kg}$ is

$$P(X_{ki_kg} = x_{ki_kg}) = \Big[p_{ki_kg} + (1 - p_{ki_kg})e^{-\mu_{ki_kg}}\Big]^{I(x_{ki_kg}=0)}$$
$$\left[(1 - p_{ki_kg})\frac{\mu_{ki_kg}^{x_{ki_kg}}}{(x_{ki_kg})!}e^{-\mu_{ki_kg}}\right]^{I(x_{ki_kg}>0)}.$$

Finally, the discriminant score $d_k(x^*)$ is,

$$\begin{aligned} d_k(\boldsymbol{x}^*) \ &= \sum_{g=1}^{G} I_{(X_g^*=0)} \log\left(\hat{p}_{kg}^* + (1 - \hat{p}_{kg}^*)e^{-d_{kg}s^*\lambda_g}\right) \\ &- \sum_{g=1}^{G} I_{(X_g^*>0)} d_{kg}s^*\lambda_g + \sum_{g=1}^{G} I_{(X_g^*>0)} \log(1 - \hat{p}_{kg}^*) \\ &+ \sum_{g=1}^{G} I_{(X_g^*>0)} X_g^* \log(d_{kg}) + \log \pi_k + C. \end{aligned} \tag{6}$$

2.5. Zero-Inflated Negative Binomial Logistic Discriminant Analysis

2.5.1. Model

In this section, we extend the zero-inflated Poisson distribution to the zero-inflated negative binomial distribution and propose ZINBLDA to model overdispersion data with excess zeros. The distribution is expressed as

$$X_{ki_kg} \sim \begin{cases} \delta_{\{0\}}, & p_{ki_kg}, \\ \text{NB}(\mu_{ki_kg}, \phi_g^{'}), & (1 - p_{ki_kg}). \end{cases}$$

Thus, the probability of X_{ki_kg} is written as

$$P(X_{ki_kg}) = \begin{cases} p_{ki_kg} + (1 - p_{ki_kg})(\frac{1}{1+\mu_{ki_kg}\phi'_g})^{\phi_g^{'-1}}, & X_{ki_kg} = 0, \\ \\ (1 - p_{ki_kg})\frac{\Gamma(X_{ki_kg}+\phi_g^{'-1})}{X_{ki_kg}!\Gamma(\phi_g^{'-1})}(\frac{\mu_{ki_kg}\phi'_g}{1+\mu_{ki_kg}\phi'_g})^{X_{ki_kg}}(\frac{1}{1+\mu_{ki_kg}\phi'_g})^{\phi_g^{'-1}}, & X_{ki_kg} > 0. \end{cases}$$

The probability density function of $X_{ki_kg} = x_{ki_kg}$ is

$$P(X_{ki_kg} = x_{ki_kg}) = \Big[(1 - p_{ki_kg})\frac{\Gamma(x_{ki_kg} + \phi_g^{'-1})}{x_{ki_kg}!\Gamma(\phi_g^{'-1})}(\frac{\mu_{ki_kg}\phi'_g}{1 + \mu_{ki_kg}\phi'_g})^{x_{ki_kg}} (\frac{1}{1 + \mu_{ki_kg}\phi'_g})^{\phi_g^{'-1}}\Big]^{I(x_{ki_kg}>0)} \quad (7)$$

$$\left[p_{ki_kg} + (1 - p_{ki_kg})(\frac{1}{1 + \mu_{ki_kg}\phi'_g})^{\phi_g^{'-1}}\right]^{I(x_{ki_kg}=0)}. \quad (8)$$

By Bayes' rule, we obtain the discriminant score $d_k(\boldsymbol{x}^*)$ of ZINBLDA using

$$\begin{aligned} d_k(\boldsymbol{x}^*) = & \sum_{g=1}^{G} I_{(X_g^*=0)} \log\Big[(1 - \hat{p}_{kg}^*)(\frac{1}{1 + s^*\lambda_g d_{kg}\phi'_g})^{\phi_g^{'-1}} \\ & + \hat{p}_{kg}^*\Big] + \sum_{g=1}^{G} I_{(X_g^*>0)} \log(1 - \hat{p}_{kg}^*) \\ & + \sum_{g=1}^{G} I_{(X_g^*>0)} X_g^* [\log d_{kg} - \log(1 + s^*\lambda_g d_{kg}\phi_g^{'})] \\ & - \sum_{g=1}^{G} I_{(X_g^*>0)} \phi_g^{'-1} \log(1 + s^*\lambda_g d_{kg}\phi_g^{'}) \\ & + \log \pi_k + C. \end{aligned} \quad (9)$$

2.5.2. Parameters Estimation

Next, we estimate the parameters in the ZINBLDA model, which includes the class difference parameter d_{kg}, size factors s_{i_k} and s^*, dispersion parameter ϕ'_g, and the probability of excess zeros p_{ki_kg}.

2.5.2.1. Class Difference Parameter Estimation

Similar to the former three methods, to estimate d_{kg} we first obtain the maximum likelihood estimation $\hat{d}_{kg} = (\sum_{i_k=1}^{n_k} X_{i_kg})/(\sum_{i_k=1}^{n_k} s_{i_k}\lambda_g)$ and then take a Gamma(β, β) prior in case of $\sum_{i_k=1}^{n_k} X_{i_kg} = 0$. Therefore, the posterior mean

$$\hat{d}_{kg} = (\sum_{i_k=1}^{n_k} X_{i_kg} + \beta)/(\sum_{i_k=1}^{n_k} s_{i_k}\lambda_g + \beta)$$

is our estimation. For convenience and due to the small influence of β on the estimation result, we assume $\beta = 1$ in this work.

2.5.2.2. Size Factor Estimation

The total number of reads between samples differs due to various sequencing depths. Generally, data must be normalized by their size factor. The three existing classification methods (PLDA, NBLDA, and ZIPLDA) use three different normalization methods: total count (Dillies et al., 2013), median ratio (Love et al., 2014), and quantile (Bullard et al., 2010). Note that there is little difference in the performance of classification among the three normalization methods. In this work, we use total count to estimate the size factor for convenience. Therefore, the estimation of size factor $\hat{s}_{i_k}$ for the training data is

$$\hat{s}_{i_k} = \frac{\sum_{g=1}^{G} X_{i_kg}}{\sum_{k=1}^{K}\sum_{i_k=1}^{n_k}\sum_{g=1}^{G} X_{i_kg}},$$

and the estimation of size factor $\hat{s}^*$ for the testing data is

$$\hat{s}^* = \frac{\sum_{g=1}^{G} X_g^*}{\sum_{k=1}^{K}\sum_{i_k=1}^{n_k}\sum_{g=1}^{G} X_{i_kg}}.$$

2.5.2.3. Dispersion Parameter Estimation

Since ZINBLDA assumes that data follow a mixture distribution rather than a negative binomial distribution, the method used to estimate the dispersion parameter in NBLDA is not applicable in this case. Therefore, we used the maximum likelihood to estimate ϕ'_g. Based on equation (8), the log likelihood function of ZINBLDA is

$$\begin{aligned} L = & \sum_{g=1}^{G}\{I_{(x_{ki_kg}=0)} \log[\hat{p}_{ki_kg} + (1 - \hat{p}_{ki_kg}))(\frac{1}{1 + \hat{\mu}_{ki_kg}\phi'_g})^{\phi_g^{'-1}}] \\ & + I_{(x_{ki_kg}>0)}[\log(1 - \hat{p}_{ki_kg}) + \log\Gamma(x_{ki_kg} + \phi_g^{'-1}) \\ & - \log\Gamma(\phi_g^{'-1}) - \log\Gamma(x_{ki_kg}!) \\ & + x_{ki_kg}\log\hat{\mu}_{ki_kg}\phi'_g - x_{ki_kg}\log(1 + \hat{\mu}_{ki_kg}\phi'_g) \\ & - \phi_g^{'-1}\log(1 + \hat{\mu}_{ki_kg}\phi'_g)]\}. \end{aligned} \quad (10)$$

Because the parameter p_{ki_kg} must also be estimated, we cannot directly take the partial derivatives and let the result equal zero to get the estimation of dispersion parameter $\hat{\phi}'_g$ in formula (10). Therefore, we first set an initial value for parameters p_{ki_kg} and ϕ_g, and then we used the PORT routines optimization method (David, 1990) to get the estimation value $\hat{\phi}'_g$.

2.5.2.4. The Probability of Excess Zeros Estimation

Assuming the data for the classifier follow a zero-inflated mixture distribution, we needed to estimate the probability of excess zeros in the distribution. Based on the process proposed by Zhou et al. (2018), we assumed that the probability of zeros, the mean of the genes, and the sequencing depth have the following logistic relation:

$$log\{\frac{P(X_{ki_kg} = 0)}{1 - P(X_{ki_kg} = 0)}\} = \alpha + \beta_1(\frac{N_{ki_k}}{N_{1i_1}}) + \beta_2\mu_{ki_kg}. \quad (11)$$

Replacing $P(X_{ki_kg} = 0)$ in model (11) with $p_{ki_kg} + (1 - p_{ki_kg})(\frac{1}{1+\mu_{ki_kg}\phi_g^{'}})^{\phi_g^{'-1}}$, we get

$$\hat{p}_{ki_kg} = \frac{p_1 - (1 + p_1)(\frac{1}{1+\mu_{ki_kg}\hat{\phi}_g^{'}})^{\hat{\phi}_g^{'-1}}}{(1 + p_1)[1 - (\frac{1}{1+\mu_{ki_kg}\hat{\phi}_g^{'}})^{\hat{\phi}_g^{'-1}}]},$$

where $p_1 = exp\{\alpha + \beta_1(\frac{N_{ki_k}}{N_{1i_1}}) + \beta_2\mu_{ki_kg}\}$; $N_{ki_k} = \sum_{g=1}^{G} X_{ki_kg}$; and $\alpha,\ \beta_1$, and β_2 are coefficients in the logistic model (11).

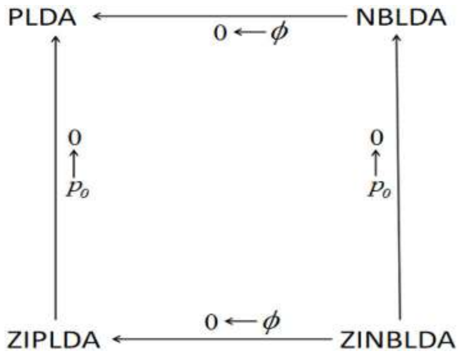

FIGURE 1 | The transformation relation between the four methods. Given a sufficient sample size, the four models can transform into one another according to the value of dispersion parameter ϕ and the average probability of excess zeros p_0.

2.6. Transformation Relation

Note that the above four models can transform into one another under some conditions.

(1) From the discriminant score function of NBLDA (formula 5), we found that if $s^*\lambda_g d_{kg}$ is bounded and $\phi_g \rightarrow 0$, then $\log(1 + s^*\lambda_g d_{kg}\phi_g) \rightarrow 0$ and $\phi_g^{-1}\log(1 + s^*\lambda_g d_{kg}) \rightarrow s^*\lambda_g d_{kg}$. Therefore, the discriminant score of NBLDA approaches that of PLDA (formula 3). That is, the NBLDA classifier reduces to the PLDA classifier when the dispersion value tends to zero.

(2) For the discriminant score function of ZIPLDA (formula 6), when $\hat{p}^*_{kg} \rightarrow 0$, then $\log(\hat{p}^*_{kg} + (1 - \hat{p}^*_{kg})e^{-d_{kg}s^*\lambda_g}) \rightarrow -d_{kg}s^*\lambda_g$, and the discriminant score of ZIPLDA approaches that of PLDA. Thus, with the probability of zeros decreased to zero, the ZIPLDA score reduces to the PLDA score.

(3) Similarly, for the discriminant score of ZINBLDA (formula 9), when $\phi'_g \rightarrow 0$, then $\phi_g^{'-1}\log(1 + s^*\lambda_g d_{kg}\phi'_g) \rightarrow d_{kg}s^*\lambda_g$ and $(1 + s^*\lambda_g d_{kg}\phi'_g)^{-\phi_g^{'-1}} \rightarrow exp\{-d_{kg}s^*\lambda_g\}$. That is, when dispersion tends to zero, the discriminant score of ZINBLDA reduces to that of the ZIPLDA. Furthermore, if $\hat{p}^*_{kg} \rightarrow 0$, then $\log[(1 - \hat{p}^*_{kg})(\frac{1}{1+s^*\lambda_g d_{kg}\phi'_g})^{\phi_g^{'-1}} + \hat{p}^*_{kg}] \rightarrow -\phi_g^{'-1}\log(1 + s^*\lambda_g d_{kg}\phi'_g)$. Therefore, when the probability of $\delta_{\{0\}}$ tends to zero, the ZINBLDA classifier reduces to the NBLDA classifier.

Figure 1 shows the above transformation relations, where ϕ denotes the dispersion parameter, and p_0 denotes the average probability of excess zeros. Starting at the bottom right of the figure and going clockwise, ZINBLDA reduces to ZIPLDA as $\phi \rightarrow 0$, and ZIPLDA reduces to PLDA as $p_0 \rightarrow 0$. Likewise, starting at the bottom right corner and going counterclockwise, ZINBLDA reduces to NBLDA as $p_0 \rightarrow 0$, and NBLDA reduces to PLDA as $\phi \rightarrow 0$. The transformation relationship between the four classification methods indicates that for data without dispersed biological replicates and excess zeros, PLDA may perform better than the other methods. However, NBLDA is good at dealing with overdispersion data, while ZIPLDA is designed to handle data with excess zeros. For data with excess zeros and dispersed biological replicates, ZINBLDA may be the optimal choice.

3. SIMULATION STUDIES

We evaluated the performance of the four methods by conducting simulations in various scenarios. We also built a decision tree to help us select the optimal classifier from the four methods for a new dataset.

3.1. Simulation Design

To ensure a fair comparison between the four classifiers, we followed the same process as Zhou et al. (2018) and generated simulation data from the following negative binomial distribution:

$$X_{ki_kg} \sim \text{NB}(d_{kg}s_{i_k}\lambda_g, \phi).$$

We set $K = 2$ to illustrate the binary classification, and each class included about $n/2$ samples. We also considered multiple classifications with $K = 3$, with each class including approximately $n/3$ samples. The rest of the distribution parameters were as follows: the size factors s_{i_k} had a uniform distribution at [0.2, 2.2], the λ_g values had an exponential distribution with an expectation of 25, and the $\log d_{kg}$ values had a normal distribution with a location of 0 and scale of σ^2 (where $\sigma = 0.2$). In the simulation studies, the *DE rate* represented the proportion of differentially expressed genes, and p was the number of genes in the samples. For simplicity, we denoted p_0 as the probability of excess zeros. In each simulation study, we changed one parameter and fixed the others, then compared the misclassification rates of the four classifiers. We specified the values for p, *DE rate*, p_0, ϕ, and n in each simulation study. Each simulation was repeated 1,000 times, and the average misclassification rates were calculated for the four methods.

3.2. Simulation Results

Study 1 investigated the impact of the dispersion parameter on the performance of the four classification methods. Considering a binary classification, we set the probability of excess zeros of data to 0 and generated 50 training and 50 testing samples. Each sample included 100 genes, 20% of which were DE genes. **Figure 2** shows the average misclassification rates of the four methods with different dispersions. Overall, the misclassification rates of the four classifiers decreased when the dispersion parameters changed from 1 to 0. PLDA and ZIPLDA showed similar performance, and both were slightly worse than NBLDA and ZINBLDA in different dispersion settings. However, when the dispersion was reduced to zero, the misclassification rates of all four methods tended to zero. From the expressions of the negative binomial and Poisson distributions, the former reduced to the latter when the dispersion parameter was reduced to zero, which indicates that NBLDA and ZINBLDA (which are based on negative binomial distribution) are more suitable for classifying overdispersion data. In addition, we changed the probability of excess zeros of simulation data from 0 to 0.1, and the other parameters remained the same. **Supplementary Figure 1**

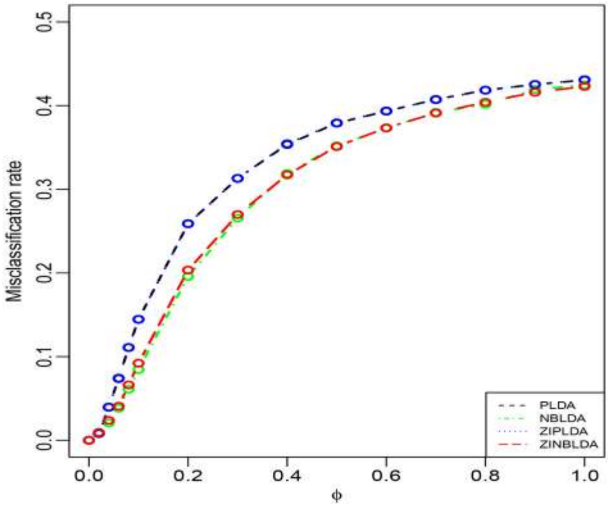

FIGURE 2 | The misclassification rates of the four methods with different dispersions (Study 1). Here, $\sigma = 0.2$, $p = 100$, $n = 50$, $p_0 = 0$, and *DE rate* $= 0.2$.

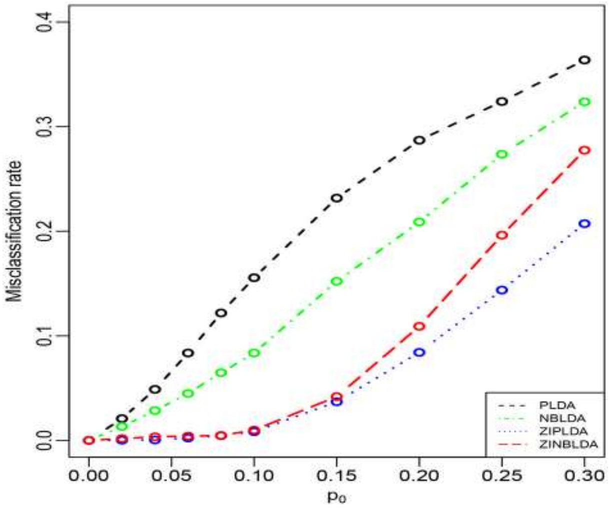

FIGURE 3 | The misclassification rates of the four methods with different probabilities of excess zeros (Study 2). Here, $\sigma = 0.2$, $p = 100$, $n = 50$, $\phi = 0$, and *DE rate* $= 0.2$.

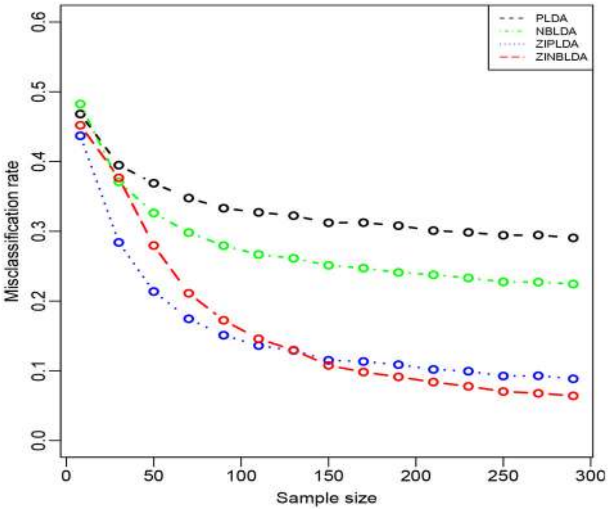

FIGURE 4 | The misclassification rates of the four methods with different sample sizes (Study 3). Here, $\sigma = 0.2$, $p = 100$, $\phi = 0$, $p_0 = 0.3$, and *DE rate* $= 0.2$.

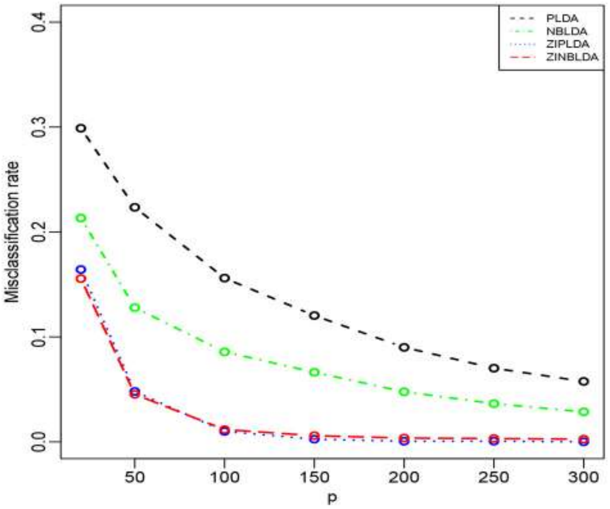

FIGURE 5 | The misclassification rates of the four methods with different numbers of genes (Study 4). Here, $\sigma = 0.2$, $n = 50$, $\phi = 0$, $p_0 = 0.1$, and *DE rate* $= 0.2$.

shows that when dispersion changed the value from 0 to 1, ZINBLDA outperformed the other methods. However, NBLDA and PLDA performed worse than ZINBLDA and ZIPLDA when the dispersion tended to zero. This result indicates that the probability of excess zeros has a major effect on the performance of the four methods.

Study 2 investigated the performance of the four methods with different probabilities of excess zeros. In this study, we set the dispersion parameter to 0, and $p_0 \in [0.1, 0.3]$; the rest of the parameters were the same as in Study 1. **Figure 3** shows that the average misclassification rates of the four classifiers increased as the probability of excess zeros increased. ZIPLDA performed slightly better than ZINBLDA when p_0 tended to 0. The performance of these two classifiers was far better than the other two classifiers, and PLDA performed the worst with different probabilities of excess zeros. This result demonstrates that ZIPLDA and ZINBLDA (which are designed for excess zeros) have a clear advantage over the other two methods when classifying data with excess zeros. Setting $\phi = 0$ could explain why ZIPLDA performed slightly better than ZINBLDA. In addition, when we reduced the sample size from 50 to 8, the result (**Supplementary Figure 2**) showed that ZIPLDA still performed the best out of the four classifiers; however, ZINBLDA performed the worst in this case, which indicates that the sample size has a major effect on the performance of ZINBLDA.

Figure 4 shows the performance of the four classification methods when the sample size changes. In Study 3, we set the probability of excess zeros to 0.3, and the sample size was gradually changed from 8 to 300. The rest of the parameters were the same as in Study 2. The overall misclassification rates gradually declined to nearly a constant value for all four classifiers when the sample size increased. ZIPLDA showed superiority over the other methods when the sample size was less than 130, and ZINBLDA attained a lower misclassification rate when the sample size was over 150. The same pattern existed between NBLDA and PLDA. When the sample size was less than 20, PLDA

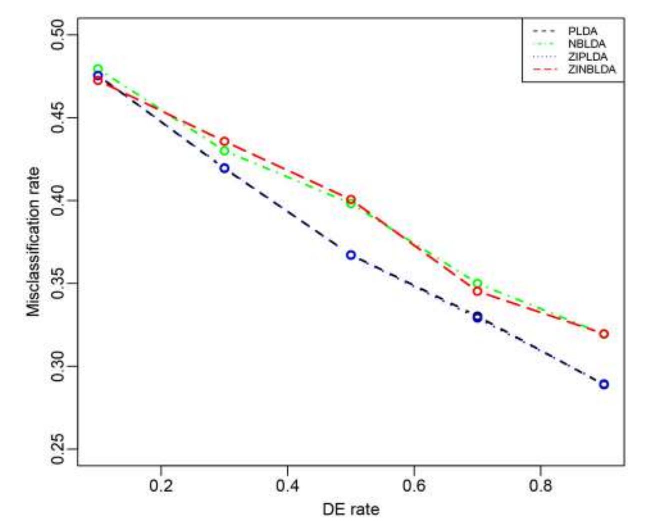

FIGURE 6 | The misclassification rates of the four methods with different probabilities of differential expression genes (Study 5). Here, $\sigma = 0.2$, $p = 100$, $n = 8$, $\phi = 0.5$, and $p_0 = 0$.

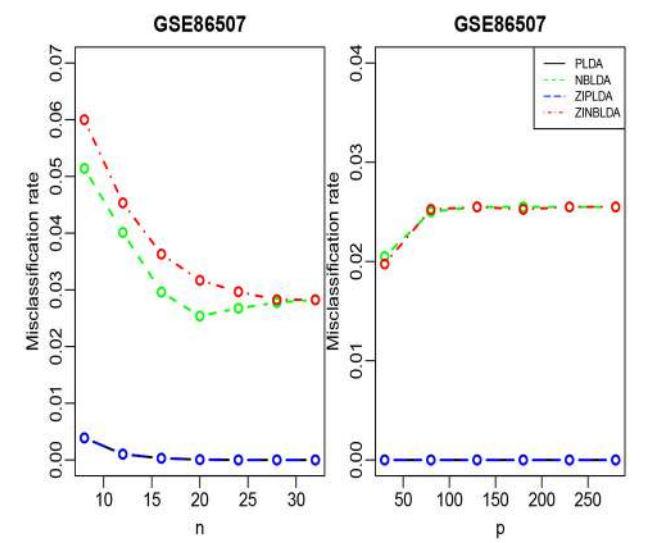

FIGURE 8 | The misclassification rates of the four classifiers for the GSE86507 dataset.

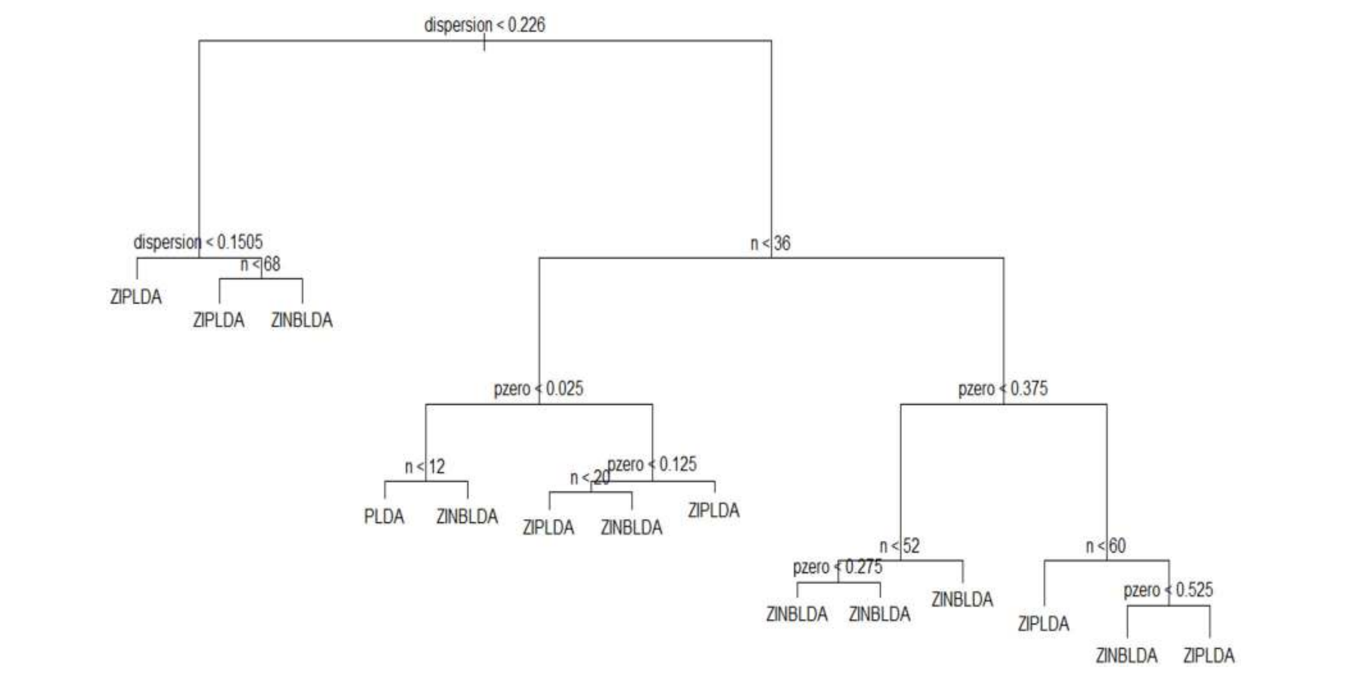

FIGURE 7 | The decision tree model used to choose an optimal classifier.

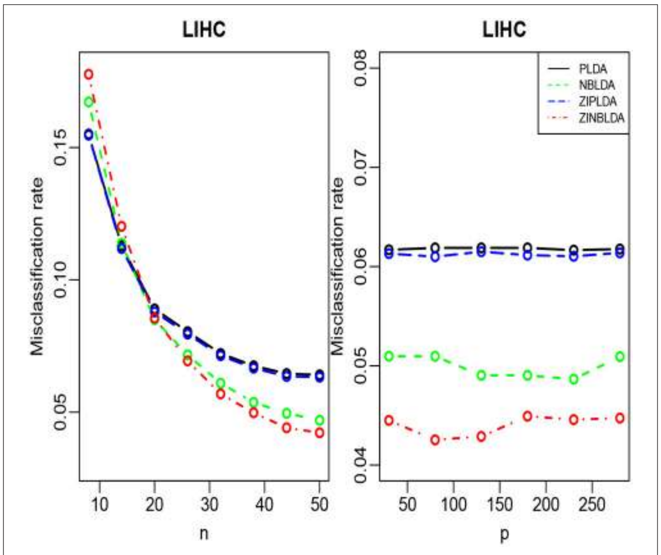

FIGURE 9 | The misclassification rates of the four classifiers for the LIHC dataset.

had a lower misclassification rate; however, NBLDA yielded a lower value when the sample size increased. The results illustrate that sample size has a huge impact on the performance of ZINBLDA and NBLDA, and ZINBLDA outperformed the other methods when a sufficient sample size was available. The reason for this may be that ZINBLDA requires a minimal number of samples to estimate the parameters in the model.

In the above three studies, we fixed the gene number at 100. In Study 4, we changed the number of selected genes and evaluated the performance of the four classifiers. The parameters were the same as in the former studies except $\phi = 0$ and $p_0 = 0.1$. **Figure 5** shows that the misclassification rates of the four methods declined as the number of genes selected increased. ZINBLDA and ZIPLDA showed similar performance and outperformed the other two methods, and PLDA performed the worst of the four methods. In **Supplementary Figure 3**, we changed the data dispersion from 0 to 0.2. A lower misclassification rate was obtained by ZINBLDA and NBLDA, and PLDA again performed the worst of the four methods. This result agrees with the conclusion that dispersion affects the performance of PLDA and ZIPLDA.

Study 5 investigated the influence of the probability of differential expression in the selected genes on the performance of the four classifiers. In this study, we set the dispersion parameter to 0.5, the probability of excess zeros was set at 0, and 100 genes were selected for all eight samples. **Figure 6** shows that the overall misclassification rates of the four methods decreased as the DE rate increased. PLDA and ZIPLDA showed similar performances, and both performed better than NBLDA and ZINBLDA with different DE rates. ZINBLDA and NBLDA performed nearly same with different probabilities of DE genes. This result demonstrates that the sample size has a marked impact on the performance of the four classifiers. In addition, **Supplementary Figure 4** shows that when the dispersion was set to 0.2 and the probability of zeros to 0.1, ZIPLDA performed remarkably better than the other methods with an increasing number of DE rates, followed by PLDA and then NBLDA, ZINBLDA performed the worst. This indicates that excess zeros in the data enable ZIPLDA to perform better than PLDA, and the sample size affects the performance of ZIPLDA notably.

We also conducted Simulation Studies 1–5 using multiple classifications ($K = 3$). **Supplementary Figures 5–9** show the performance of the four methods. The parameters were the same as in Studies 1–5 except for the sample sizes. We set $n = 75$ in **Supplementary Figures 5**, **6**, **8**; $n = 12$ in **Supplementary Figure 9**; and $n = 450$ in **Supplementary Figure 7** and compared the results with those of Studies 1–5. The performance of the four classifiers remained the same as in the binary classification.

In the simulation studies conducted above, the performance of the four classifiers was related to the attributes of the dataset, including sample size n, dispersion parameter ϕ, and the probability of excess zeros p_0. In the final simulation study, we considered a binary classification with three changeable parameters and compared the performance of the four methods for different combinations of those three parameters. We still selected 100 genes, 40% of which were DE genes. The probability of excess zeros was set at 0.001, 0.1, or 0.3, and the sample size was 8, 50, or 100. The dispersion parameters changed from 0.001 to 0.1 to 1 with every 0.2 steps. The average misclassification rates of the four methods are shown in **Supplementary Figure 10**. To clarify display the result, **Supplementary Table 1** shows the concrete values of each misclassification rate. Comparing the results of the three panels in each column, we found that for the first column (sample size of 8), the overall misclassification rates of the four methods increased when the probability of excess zeros increased from 0.001 to 0.3. When the probability of excess zeros was equal to 0.3, the misclassification rates approached 50%. When $p_0 = 0.001$, the performance of ZIPLDA and PLDA was better than NBLDA and ZINBLDA. However, when $p_0 = 0.1$, ZIPLDA outperformed the other methods, which indicates that ZIPLDA is more suitable for handling data with a small probability of excess zeros, and the sample size has less of an impact on it. When the sample size was increased to 50 (the second column), the overall performance of ZINBLDA was slightly better than that of ZIPLDA except when ϕ was small and $p_0 = 0.3$. The reason for this may be that there were not enough samples to estimate the parameters of ZINBLDA. Therefore, when the sample size increased to 100 (the third column), that ZINBLDA yielded a lower or equal misclassification rate compared to the other methods, which indicates that ZINBLDA can achieve the best classification result as long as enough samples are available. The performance of ZIPLDA also improved when p_0 increased from 0.1 to 0.3 due to the increase in the probability of excess zeros.

3.3. Optimal Classifier Selection

To select an optimal classification method for different datasets, we built a decision tree and a random forest. A decision tree is a machine learning algorithm that is widely used in many scenarios because of its accuracy for the current algorithms. As its name implies, a decision tree is a decision support tool that uses a tree-like model. It is comprised of nodes and branches, and each sample is tested on an internal node. The outcome of the test determines which branch is followed, and this procedure continues until the leaf node that holds the class label of this sample is reached. Random forest is an ensemble of a decision tree, and it can achieve a more stable result than a decision tree.

To employ a tree-like model to select the optimal classifier, the chosen features were the sample number n, dispersion ϕ, and probability of excess zeros p_0, and y was regarded as the optimal classification method. The parameter region was divided to assign the value of the feature vector. The values of sample size n ranged from 8 to 100 with a step size of 8. The dispersion parameter ranged from 0.001 to 1.001 at intervals of 0.1. The probability of excess zeros p_0 ranged from 0 to 0.6 with a step size of 0.05. For each calculation, we took one value from each parameter set to generate the simulation data, allowing for multiple combinations of these three parameters. This procedure was repeated 1,000 times, and the classifier corresponding to the smallest value of the average misclassification rate was regarded as the optimal classification method. We used the obtained data to train a decision tree, and **Figure 7** displays the classification result. This model fits the data very well, with a misclassification rate of only 7.4%. To use this model, we only need to know or estimate the values of the three parameters, then use the conditional control statements in the decision tree to distinguish in each internal node, which will result in the optimal method when the leaf node is reached. In this way, this model can be used to help choose the optimal classification method. Similarly, we can obtain a random forest with a lower misclassification rate (2.2%). The classification results of decision tree and random forest are saved in R scripts, which could be used to choose the optimal classifier when inputting the parameters of dataset.

4. APPLICATION TO REAL DATA

We further compared the four methods by analyzing two real datasets: GSE86507 and TCGA-LIHC (Liver Hepatocellular Carcinoma). The details of these two RNA-seq datasets are as follows.

Woo et al. (2017) created the GSE86507 dataset to compare gene expression between two mouse models, Pkd1f/f: HoxB7-cre mice and Pkd2f/f: HoxB7-cre mice. Each group includes 18 samples, and there are a total of 29,996 transcripts in this dataset. It contains about 17.74% zeros of all numerical values.

The dataset TCGA-LIHC contains two groups of samples: the normal group (340 samples) and the cancerous group (50 samples). There are 60,487 genes in this dataset, which contains about 43.24% zeros of all numerical values.

We chose to classify parts of genes since the majority of genes in a dataset are not differentially expressed and thus do not contribute to the sample classification. Including entire genes in the model would reduce the classification accuracy and increase the computational complexity. Thus, selecting parts of genes not only improves the accuracy of classification but saves computation time. Following the steps outlined by Dudoit et al. (2002), we selected genes by first calculating the ratio of the sum of the squares between groups and within groups for each gene, then sorted all of the genes according to the ratio from greatest to least, and finally selected a certain number of genes for downstream analysis.

We randomly split the data into a training set and test set, with both datasets containing all classes. We selected the 300 most differentially expressed genes to train the model. This procedure was repeated 1,000 times, and the average misclassification rates for each method were recorded. The left panels of **Figures 8**, **9** show that for the test data, the average misclassification rates of the four methods decreased as the number of training data gradually increased. For the GSE86507 dataset, the misclassification rates of PLDA and ZIPLDA were lower than NBLDA and ZINBLDA, both of which were close to zero. However, for the TCGA-LIHC dataset, PLDA and ZIPLDA were superior to NBLDA and ZINBLDA when the sample size was small. As the training sample size increased, the misclassification rates of NBLDA and ZINBLDA decreased remarkably, and ZINBLDA outperformed the other three methods for a large sample size. We also evaluated the classification performance of the four methods by fixing 30 training sets and gradually increasing the number of selected DE genes. The right panel of **Figure 8** shows that PLDA and ZIPLDA outperformed the other two methods, whereas the right panel of **Figure 9** shows the superiority of ZINBLDA over the other methods in this case.

To assess the efficiency of the decision tree model, we estimated the dispersion and probability of excess zeros for the two datasets. The estimated dispersion of GSE86507 was $\phi = 0.12$, and the probability of excess zeros was 0.5%, which indicates that the dataset has slight overdispersion and almost no excess zeros. The estimated dispersion of TCGA-LIHC was $\phi = 1.12$, and the probability of excess zeros was 8%, which indicates that the dataset has high overdispersion and many excess zeros. According to the conclusions in section 2.6, PLDA should perform better with the GSE86507 dataset, and ZINBLDA should be the optimal method to classify the TCGA-LIHC dataset. We used the estimated parameters to select the optimal method according to the conditional control statements in the decision tree model (**Figure 7**). Based on the result, we recommend selecting ZIPLDA for the GSE86507 dataset and ZINBLDA for the TCGA-LIHC dataset, which coincides with the real analysis results.

5. DISCUSSION

RNA-seq data classification is vital to the diagnosis of diseases. In this work, we extended the existing classification methods and proposed a ZINBLDA method for overdispersion RNA-seq data with an excess of zeros. Concretely, we built a mixture distribution with a point mass at zero and a negative distribution to model the data, and a logistic regression was used to build a relation between the probability of zeros, the mean of the genes, and the sequencing depth. Most importantly, we examined four classification methods from the perspective of their parameters, and we found that these four methods can transform into each other in some cases.

In the simulation studies, we evaluated the performance of the four methods in a wide range of settings. The simulation results showed that different methods perform better for different applications. In addition, we found that the application region of each method is associated with the attributes of the dataset, such as the dispersion, sample size, and probability of excess zeros. Therefore, we built a decision tree to help us select the optimal classification methods in different cases. In the real data analysis, we analyzed two real, next-generation sequencing datasets, and the results further confirmed the theory and simulation conclusions.

Although each of the four methods performed well in certain scenarios, there are numerous issues that remain to be solved, such as single cell RNA-seq data being particularly prone to dropout events due to the relatively shallow sequencing depth per cell. In this case, the existing classification methods may not provide a good result in practice. Therefore, we plan to develop a new classification method that employs deep learning technology to model scRNA-seq data to further improve our current work.

DATA AVAILABILITY STATEMENT

The dataset GSE86507 for this study can be found in the NCBI repository (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE86507), and the dataset TCGA-LIHC for this study can be found in the GDC repository (https://portal.gdc.cancer.gov/projects/TCGA-LIHC).

AUTHOR CONTRIBUTIONS

YZ conceived the idea. JZ processed the data and conducted simulation and real dataset experiments. JZ, ZY, and YZ wrote the manuscript. LS, MZ, YZ, and WL revised the manuscript. All authors read and approved the final manuscript.

FUNDING

YZ's research was supported by the National Natural Science Foundation of China (Grant Nos. 12071305, 11871390, and 11871411), Natural Science Foundation of Guangdong Province of China under grant 2020B1515310008, Project of Educational Commission of Guangdong Province of China under grant 2019KZDZX1007. MZ's research was supported by the Social Science Foundation of China (Grant No. 15CTJ008). This paper is partially supported by the National Natural Science Foundation of China (No. 11901401).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.642227/full#supplementary-material

REFERENCES

Berger, M., Levin, J., Vijayendran, K., Sivachenko, A., Adiconis, X., Maguire, J., et al. (2010). Integrative analysis of the melanoma transcriptome. *Genome Res.* 20, 413–427. doi: 10.1101/gr.103697.109

Biesecker, L., Burke, W., Kohane, I., Plon, S., and Zimmern, R. (2012). Next-generation sequencing in the clinic: are we ready? *Nat. Rev. Genet.* 13, 818–824. doi: 10.1038/nrg3357

Bullard, J., Purdom, E., Hansen, K., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics* 11:94. doi: 10.1186/1471-2105-11-94

David, M. G. (1990). *Usage Summary for Selected Optimization Routines*. Computing Science Technical Report 153. AT T Bell Laboratories, Murray Hill, NY.

Dillies, M., Rau, A., Aubert, J., Christelle, H., Marine, J., Nicolas, S., et al. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* 14, 671–683. doi: 10.1093/bib/bbs046

Dong, K., Zhao, H., Tong, T., and Wan, X. (2016). NBLDA: negative binomial linear discriminant analysis for RNA-Seq data. *BMC Bioinformatics* 17:369. doi: 10.1186/s12859-016-1208-1

Dudoit, S., Fridlyand, J., and Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* 97:77–87. doi: 10.1198/016214502753479248

Lin, B. Q., Zhang, L. F., and Chen, X. (2014). LFCseq: a nonparametric approach for differential expression analysis of RNA-seq data. *BMC Genomics* 15:S7. doi: 10.1186/1471-2164-15-S10-S7

Love, M., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* 9, 387–402. doi: 10.1146/annurev.genom.9.081307.164359

Morozova, O., Hirst, M., and Marra, M. A. (2009). Applications of new sequencing technologies for transcriptome analysis. *Annu. Rev. Genomics Hum. Genet.* 10, 135–151. doi: 10.1146/annurev-genom-082908-145957

Nagalakshmi, U., Zhong, W., Karl, W., Chong, S., Debasish, R., Mark, G., et al. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320, 1344–1349. doi: 10.1126/science.1158441

Robinson, M., McCarthy, D., and Smyth, G. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Robinson, M., and Smyth, G. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* 9, 321–332. doi: 10.1093/biostatistics/kxm030

Tan, K., Petersen, A., and Witten, D. (2014). "Classification of RNA-seq data," in *Statistical Analysis of Next Generation Sequencing Data*, eds S. Datta and D. Nettleton (New York, NY: Springer), 219–246. doi: 10.1007/978-3-319-07212-8_11

Trapnell, C., Williams, B., Pertea, G., Mortazavi, A., Kwan, G., van, B. M., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi: 10.1038/nrg2484

Witten, D. (2011). Classification and clustering of sequencing data using a Poisson model. *Ann. Appl. Stat.* 5, 2493–2518. doi: 10.1214/11-AOAS493

Woo, Y., Kim, D., Koo, N., Kim, Y., Lee, S., Ko, J., et al. (2017). Profiling of mirnas and target genes related to cystogenesis in adpkd mouse models. *Sci. Rep.* 7:14151. doi: 10.1038/s41598-017-14083-8

Zhou, Y., Wan, X., Zhang, B., and Tong, T. (2018). Classifying next-generation sequencing data using a zero-inflated Poisson model. *Bioinformatics* 34, 1329–1335. doi: 10.1093/bioinformatics/btx768

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.