Technology and Health Care 28 (2020) S161–S171
DOI 10.3233/THC-209016
IOS Press

Penalized logistic regression based on $L_{1/2}$ penalty for high-dimensional DNA methylation data

Hong-Kun Jiang* and Yong Liang
Faculty of Information Technology, Macau University of Science and Technology, Avenida Wai Long, Taipa, Macau, China

Abstract.
BACKGROUND: DNA methylation is a molecular modification of DNA that is vital and occurs in gene expression. In cancer tissues, the 5'–C–phosphate–G–3'(CpG) rich regions are abnormally hypermethylated or hypomethylated. Therefore, it is useful to find out the diseased CpG sites by employing specific methods. CpG sites are highly correlated with each other within the same gene or the same CpG island.
OBJECTIVE: Based on this group effect, we proposed an efficient and accurate method for selecting pathogenic CpG sites.
METHODS: Our method aimed to combine a $L_{1/2}$ regularized solver and a central node fully connected network to penalize group constrained logistic regression model. Consequently, both sparsity and group effect were brought in with respect to the correlated regression coefficients.
RESULTS: Extensive simulation studies were used to compare our proposed approach with existing mainstream regularization in respect of classification accuracy and stability. The simulation results show that a greater predictive accuracy was attained in comparison to previous methods. Furthermore, our method was applied to over 20000 CpG sites and verified using the ovarian cancer data generated from Illumina Infinium HumanMethylation 27K Beadchip. In the result of the real dataset, not only the indicators of predictive accuracy are higher than the previous methods, but also more CpG sites containing genes are confirmed pathogenic. Additionally, the total number of CpG sites chosen is less than other methods and the results show higher accuracy rates in comparison to other methods in simulation and DNA methylation data.
CONCLUSION: The proposed method offers an advanced tool to researchers in DNA methylation and can be a powerful tool for recognizing pathogenic CpG sites.

Keywords: DNA methylation, CpG island, $L_{1/2}$ regularization method, gene regulatory network, variable selection

1. Introduction

DNA methylations occur at cytosine which might affect the modifications of DNA molecules. In this process, the gene expressions can be regulated without changing the DNA sequences. In particular, the related gene silencing of DNA methylations is a well-accepted epigenetic mechanism that often occurs at tumor suppressor genes loci in human cancers [1–5]. Recently, some high-throughput DNA methylation platforms have generated amounts of DNA methylation data and mostly based on genotyping bisulfite converted DNA. In this paper, one of the popular platforms, Illumina Infinium HumanMethylation 27K

*Corresponding author: Hong-Kun Jiang, Faculty of Information Technology, Macau University of Science and Technology, Avenida Wai Long, Taipa, Macau 999078, China. E-mail: jiang.hongkun@qq.com.

array, was used. Additionally, the β-values indicate the methylation status of the CpG sites within the array while each site's value is calculated by the average of approximately 30 replicates [6]. Every individual β-value is a continuous variable between 0 and 1, where zero means unmethylated and one means methylated.

To date, researchers have selected methylated sites by statistical classification approaches [7–9]. Even though most of the CpG sites display various degrees of methylation, only a few gene expressions change. The statistical approaches therefore are difficult to find relevant CpG sites from high-dimensional data, making the statistical approaches not suitable for methylation data. In order to select CpG sites, different parameter models were utilized by researchers to represent diverse status of the samples [10]. Methylation data expresses different features from gene expression data. Firstly, the DNA methylation data has a group effect feature among CpG sites based on gene groups and CpG island groups. Secondly, the DNA methylation data values range between 0 and 1. Based on these features, Sun [11] has proposed a procedure that merged the L_1 penalty and squared L_2 penalty to select methylated CpG sites.

With the Illumina HumanMethylation 27K array, each gene has about 1–25 correlated CpG sites and each CpG island has about 2–11 CpG sites. Based on these aspects of DNA methylation data, a $L_{1/2}$ penalized logistic regression model has been introduced to select potentially diseased pathogenic CpG sites within one gene. The $L_{1/2}$ regularization can be represented by L_q ($0 < q < 1$) regularization and has exhibited properties for instance unbiasedness, sparsity and oracle [12]. Additionally, the sparsity of the $L_{1/2}$ regularization is better than L_1 regularization [12–14]. Based on the $L_{1/2}$ penalized logistic regression model, we used the proposed network structure (the central node fully connected network) to describe the two correlated CpG sites' patterns, one is based on the gene group, whilst the other is based on the CpG island group. The proposed method is designed to select CpG sites by group effect that associate with diseases. The aimed method has a finer specificity than present methods, as it for instance has the potential to select more relevant genes.

2. Methods

2.1. Network-regularization

In this research, n samples were used, $D = \{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)$ where $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is the methylation β-*value* of the i-th sample and p represents the total number of CpG sites, the dependent variable y_i is a binary variable where 0 implies controls and 1 implies cases. The logistic regression is:

$$f\left(x_i, \varphi\right) = \frac{exp\left(x_i^T \varphi\right)}{1 + exp\left(x_i^T \varphi\right)} \tag{1}$$

where φ is the regression coefficients. The logistic log-likelihood is defined as:

$$l\left(\varphi\right) = -n^{-1} \sum_{i=1}^{n} \left[y_i log f(x_i, \varphi) + (1 - y_i) log(1 - f(x_i, \varphi))\right] \tag{2}$$

φ was obtained by minimizing the log-likelihood. In high dimensional application, it is not appropriate to solve the logistic model directly and may result in overfitting. Hence, the regularization approaches are employed to aim at the overfitting problem. The sparse logistic regression can be laid out as Eq. (2) when a regularization term is added:

$$\varphi^* = avgmin(-l\left(\varphi\right) + P(\varphi)) \tag{3}$$

where $P(\varphi)$ is the penalty function.

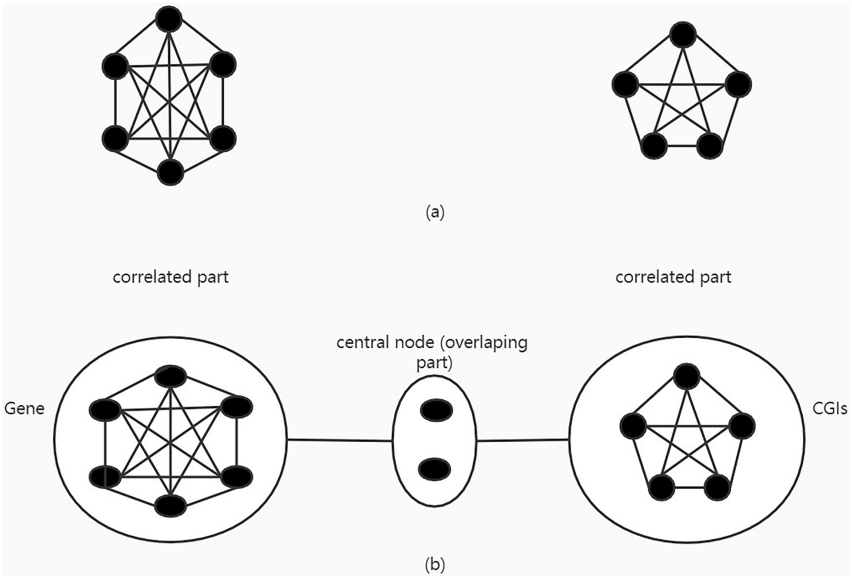

Fig. 1. a. Previous fully connected network. b. Central node fully connected network.

Lasso (L_1) and L_2, a well-recognized regularization approach was used in previous methods. The L_2 does not have sparsity and L_1 has a sparsity less than L_q ($0 < q < 1$). Nonetheless, when q lies closer to zero, results show a sparser L_q and subsequently more challenging to converge. Therefore, some researchers [12] investigated the properties of L_q ($0 < q < 1$) regularization and demonstrated the $L_{1/2}$ regularization is particularly essential and crucial. The performance between L_q penalty and $L_{1/2}$ has no significant diversity whereas the $L_{1/2}$ regularization is much more facile to solve. Accordingly, the $L_{1/2}$ regularization can be laid out as L_q ($0 < q < 1$) regularization which exhibits unbiasedness as well as oracle properties [12–14]. In high-dimensional DNA methylation data, disease-related CpG sites are very limited and therefore, in practice, the $L_{1/2}$ penalty methodology would be more significant than the L_0, L_1 and L_2 approaches. Consequently, the $L_{1/2}$ penalty was favored in our logistic regression model.

Some methods have been provided in order to tackle highly correlated variables. *Elastic net* penalty ($L_1 + L_2$) and *HLR* ($L_{1/2} + L_2$) emphasizes a grouping effect and tend to smooth the coefficient profiles. However, the pathway information was neglected in these methods. To merge CpG sites deduction into the analysis of high-demensional methylation data, we extended a network-based regularization technique designed for the $L_{1/2}$ penalty.

The methylation data displays a strong group effect and thus previous research used a fully connected network (Fc.net) to describe the correlated CpG sites group patterns within a gene. In methylation data, the group effect of CpG sites is not only present within one gene and present within one CpG island. There are overlapping parts between groups and these overlapping parts correlate with both parts respectively. With the different previous network, we set the overlapping part as the central node and connect it with other correlated parts (Fig. 1). The network not only has the genome information or CpG island information, but also the two aspects of information integrated into the network. It can better reflect the relevance of CpG site.

The network information is represented in a graphed structure with p-dimensional Laplacian matrix

$L = \{l_{ab}\}$. It is defined as:

$$l^*_{ab} = \begin{cases} 1 & \textit{if } a = b \textit{ and } d_a \neq 0 \\ -(d_a d_b)^{-1/2} & \textit{if a and b are linked with each other} \\ 0 & \textit{otherwise} \end{cases}$$

where d_a is the total number of connections at vertex a in graph.

The penalty function in Eq. (3) is:

$$\begin{aligned} P(\varphi) &= \lambda_1 ||\varphi||^{-\frac{1}{2}} + \lambda_2 \varphi^T L \varphi \\ &= \lambda_1 \sum_{i=1}^{p} |\varphi_i|^{-\frac{1}{2}} + \lambda_2 \sum_{a=1}^{p} \sum_{a \sim b} \left(\frac{\varphi_a}{\sqrt{d_a}} - \frac{\varphi_b}{\sqrt{d_b}} \right) \end{aligned} \tag{4}$$

where $||\cdot||^{1/2}$ is a $L_{1/2}$ norm and $a \sim b$ illustrate the variables which are linked to the *a-th* predictor. The sparsity and smoothness are controlled by the parameters λ_1 and λ_2.

The effectiveness of the penalty function reduced significantly when two negatively correlated predictors are interacted; the signs of coefficients are thus predicted and added to the Laplacian matrix to overcome problem:

$$l^*_{ab} = \begin{cases} 1 & \textit{if } a = b \textit{ and } d_a \neq 0 \\ -sgn(\varphi^*_a) sgn(\varphi^*_b) (d_a d_b)^{-1/2} & \textit{if a and b are linked with each other} \\ 0 & \textit{otherwise} \end{cases} \tag{5}$$

The adaptive net function can be written as:

$$\varphi^T L^* \varphi = \sum_{a=1}^{p} \sum_{a \sim b} \left(\frac{sgn(\varphi^*_a)\varphi_a}{\sqrt{d_a}} - \frac{sgn(\varphi^*_b)\varphi_b}{\sqrt{d_b}} \right)^2$$

Based on $|\beta_a| \approx sgn(\beta^*_a)\beta_a$ for $\beta_a \approx \beta^*_a$, the adaptive penalty function can be written as:

$$P(\varphi) = \lambda_1 \sum_{j=1}^{p} |\varphi_j|^{1/2} + \lambda_2 \sum_{a=1}^{p} \sum_{a \sim b} \left(\frac{|\varphi_a|}{\sqrt{d_a}} - \frac{|\varphi_b|}{\sqrt{d_b}} \right)^2 \tag{6}$$

2.2. The coordinate descent algorithm

To solve regularization models, the coordinate descent algorithm adopted as a competent tool. Regarding the coordinate descent algorithm, we referred to previous research [11,15,16] and Eq. (2) can be linearized by Taylor series expansion at current estimates φ^*:

$$l^*(\varphi) \approx \frac{1}{2} n^{-1} \sum_{i=1}^{n} w_i \left(z_i - x_i^T \varphi \right)^2 \tag{7}$$

where $z_i = x_i^T \varphi^* + (y_i - f^*(x_i))/w_i$, $w_i = f^*(x_i)(1 - f^*(x_i))$, $f^*(x_i) = exp\left(x_i^T \varphi^*\right)/(1 + exp(x_i^T \varphi^*))$.

Next, the estimator:

$$\varphi^*_a = \frac{s(n^{-1} \sum_{i=1}^{n} w_i x_{ia} \left(z_i - \tilde{z}_i^{(a)} \right) + \lambda_2 g(a), \lambda_1)}{\lambda_2 + 1} \tag{8}$$

where $\tilde{z}_i^{(a)} = \sum_{j \neq i} x_{ij}\varphi_j^*$, $g\,(a) = \sum_{a \sim b} \frac{|\varphi_b^*|}{\sqrt{d_a d_b}}$ and $s(\sigma, \gamma)$ is an enhanced $L_{1/2}$ thresholding operator for the coordinate descent algorithm [12–14].

$$S\,(\sigma, \gamma) = \begin{cases} \frac{2}{3}\sigma\left(1 + cos\left(\frac{2(\pi - \phi_\gamma(\sigma))}{3}\right)\right) & \textit{if } |\sigma| > \frac{\sqrt[3]{54}}{4}(\gamma)^{\frac{2}{3}} \\ 0 & \textit{otherwise} \end{cases} \tag{9}$$

where $\phi_\gamma\,(\sigma) = \textit{across}((\gamma/8)(|\sigma|/3)^{2/3})$.

3. Results and discussion

3.1. Analyses of simulated data

The performance of the proposed simulation study quoting the simulation from Teschendorff et al. [17] and Su and Wang [11] was analyzed and evaluated. There were 600 groups, which were divided into 100 groups, 150 groups and 7 sets of 50 groups in accordance to their number of CpG sites. Each group comprised of at least 1 CpG site up to 9 CpG sites reciprocally. In total, there were 2500 CpG sites.

First, we simulated variables with the group effect ranging between 0 and 1. So we performed an inverse logit transformation on a multivariate normal distribution variable to represent the β-*values* of the *i-th* CpG site in the *g-th* group.

$$x_{i,g} = \frac{\exp\,(t_{i,g})}{1 + \exp\,(t_{i,g})},\ t_{i,g} \sim 2N_{s_g}(\mu,\sigma) \tag{10}$$

where s_g is the size of group, i.e. $1 \leqslant s_g \leqslant 9$. In this simulation model, we set $\mu = (-1, \ldots, -1)^T$, $x_{i,g}$ ranging between 0 (unmethylated) to 1 (completely methylated). The relationship of CpG sites within group is shown by σ. The covariance matrix σ is presented as follows:

(1) $\sigma = \rho^{|a-b|}$, $\rho = 0.2, 0.5, 0.7$.
(2) $\sigma = \rho$ for a $\neq$ b and and σ =1 for $a = b$, $\rho = 0.2, 0.5, 0.7$.

The first condition is autoregressive (AR) model, and the second condition is compound symmetric correlation model. We set three different correlation coefficients $\rho = 0.2$, 0.5 and 0.7 for all conditions [18, 19].

Second, given the regression coefficients φ based on previous research, $\varphi_g = (\varphi_{1,g}, \ldots, \varphi_{p_g,g})^T$ is the coefficient of CpG sites within the *g-th* group. After that, one group from each of the 9 different groups was selected to set the regression coefficients. At this step, there were 45 CpG sites which have been assigned the regression coefficients value. The regression coefficients $\varphi_{k,g}$ are presented as:

$$\varphi_{k,g} = \begin{cases} {s_g}^{-0.5}(-1)^k\delta, & \text{for all } k = 1, \ldots, s_g \ \ \textit{if } s_g \textit{ is even number} \\ {s_g}^{-0.5}\delta, & \text{for all } k = 1, \ldots, s_g \ \ \textit{if } s_g \textit{ is odd number} \end{cases} \tag{11}$$

when δ is the strength of the true signals. The other sets of regression coefficients were set to 0.

In the simulation models, there were 45 pathogenic CpG sites in a total of 2500 CpG sites. Lastly, the y_i is given by Bernoulli distribution. For each simulation set, there were 200 cases and 200 controls. There were nine simulation conditions based on different parameters, for instance the strength of the true signals.

We repeated simulations 100 times for each condition. We then used the 10-fold cross-validation (CV) approach in the training set in order to tune the optimal regularization parameters of the *Lasso*, *Elastic-Net* (*Enet*), L_1+ *Fc.net*, $L_{1/2}$, *HLR* ($L_{1/2} + L_2$), $L_{1/2}$+ *Fc.net*. Note that, the *Enet*, L_1+ *Fc.net*,

Table 1
The total area under the averaged ROC curves (AUC) and MSE for all models

δ	σ	ρ	Lasso		Enet		L_1+ *Fc.net*		$L_{1/2}$		*HLR* ($L_{1/2}+L_2$)		$L_{1/2}$+ *Fc.net*	
			AUC	MSE	AUC	MSE	AUC	MSE	AUC	MSE	AUC	MSE	AUC	MSE
1	AR(1)	0.2	0.806	0.273	0.847	0.238	0.850	0.239	0.809	0.260	0.853	0.254	**0.860**	**0.197**
1	AR(1)	0.5	0.871	0.213	0.885	0.206	0.888	0.197	0.856	0.225	0.933	0.161	**0.954**	**0.116**
1	AR(1)	0.7	0.898	0.187	0.906	0.185	0.909	0.175	0.917	0.169	0.942	0.149	**0.962**	**0.101**
2	AR(1)	0.2	0.806	0.273	0.860	0.237	0.866	0.220	0.809	0.260	0.869	0.213	**0.921**	**0.155**
2	AR(1)	0.5	0.871	0.273	0.889	0.237	0.903	0.220	0.918	0.167	0.953	0.134	**0.975**	**0.085**
2	AR(1)	0.7	0.898	0.187	0.904	0.177	0.912	0.168	0.917	0.169	0.953	0.133	**0.970**	**0.089**
2	CS	0.2	0.852	0.226	0.879	0.207	0.889	0.193	0.820	0.257	0.893	0.197	**0.936**	**0.139**
2	CS	0.5	0.899	0.181	0.913	0.163	0.919	0.157	0.895	0.182	0.961	0.128	**0.969**	**0.097**
2	CS	0.7	0.924	0.162	0.927	0.157	0.934	0.147	0.915	0.171	0.957	0.125	**0.979**	**0.080**

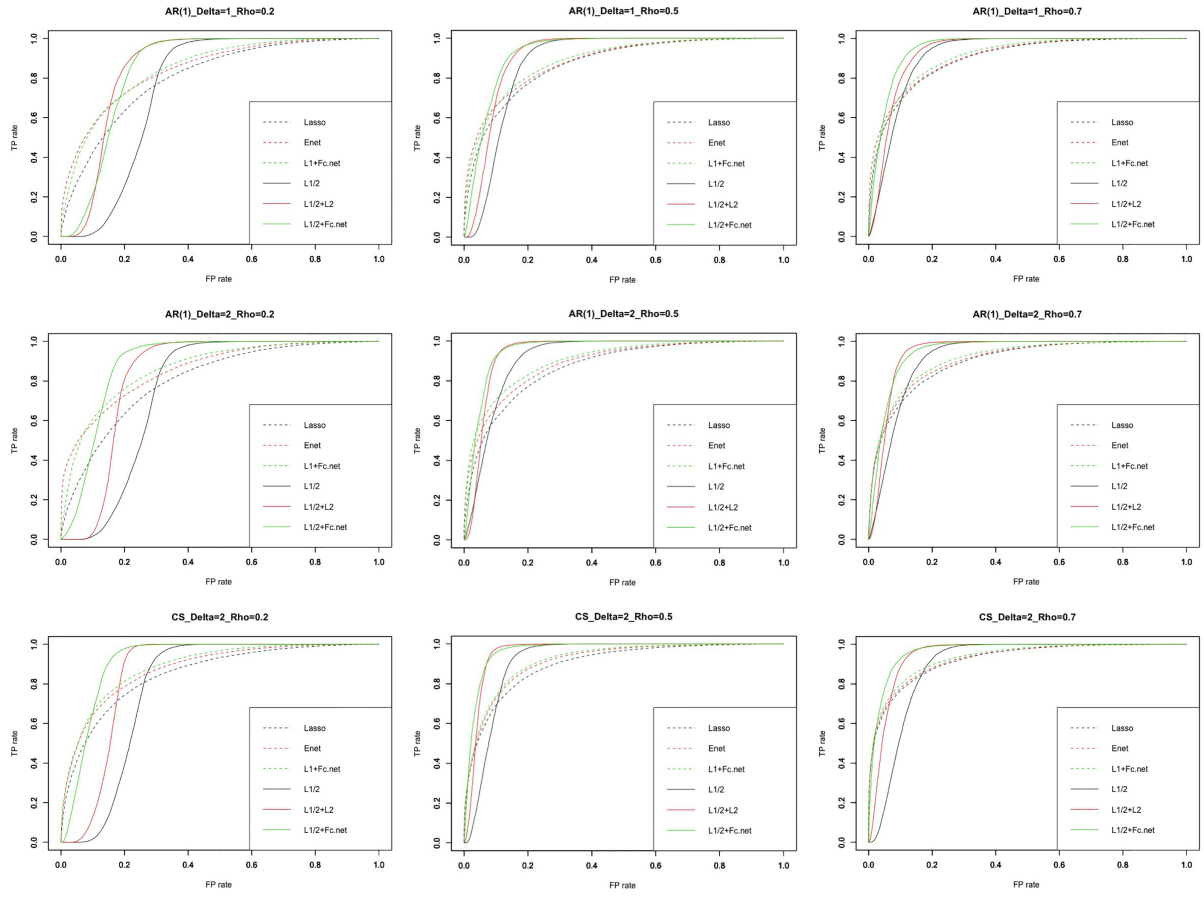

Fig. 2. The ROC curve of every model.

HLR ($L_{1/2}+L_2$) and $L_{1/2}$+ *Fc.net* methods have two-dimensional parameter surfaces in the 10-*CV* approach. Afterwards, the logistic regressions with the estimated tuning parameters were employed to build different classifiers. Lastly, the attained classifiers were adopted to the test set for further classification and prediction.

Figure 2 shows the receiver operating characteristic curve (ROC curve) for every model. The green

Table 2
The AUC of real data for each method

	Lasso	Enet	L_1+ *Fc.net(gene)*	$L_{1/2}$	*HLR* ($L_{1/2}+L_2$)	$L_{1/2}$+ *central node Fc.net*
Pre	0.798	0.886	0.921	0.803	0.908	0.946
Post	0.762	0.898	0.934	0.771	0.923	0.948

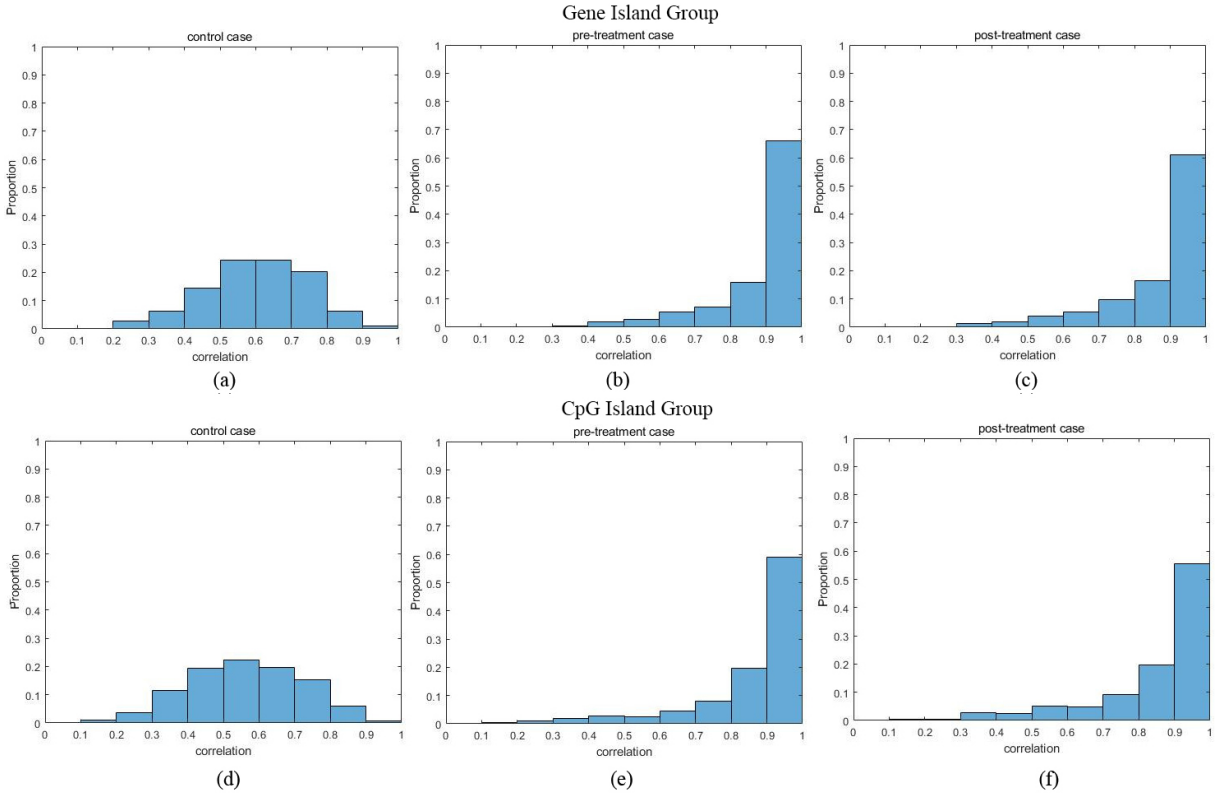

Fig. 3. The histogram of correlation between CpG sites.

solid line ($L_{1/2}$+ *Fc.net*) is closer to the upper left corner in the system than other line. So the effect of $L_{1/2}$+ *Fc.net* is at optimal for the other algorithm in each model. Table 1 shows the total area under the averaged ROC curves and the MSE of every model respectively. From Table 1 it can be seen that $L_{1/2}$+ *Fc.net* also has a very good performance within all models. In general, our proposed enhanced $L_{1/2}$ model achieved preponderant accuracy rates in all models in comparison to the other methods (the *Lasso*, *Enet*, L_1+ *Fc.net*, $L_{1/2}$ and *HLR* ($L_{1/2}+L_2$)).

3.2. Analyses of real data

To further evaluate the effectiveness of our proposed method, in this section, we examined the DNA methylation (ovarian cancer) data generated from Illumina Infiniumm HumanMethylation 27K Beadchip [20]. The data is accessible from NCBI (http://www.ncbi.nlm.nih.gov/).

The data was generated by llumina Infiniumm HumanMethylation 27K Beadchip that contains 22727 CpG sites. We first removed samples which were low in BS conversion efficiency or low in CpG coverage. After that, a total of 207 genes contained more than 3 CpG sites and 295 CpG islands contained more than

Table 3
The top 20 CpG sites and the corresponding genes selected from the comparison between pre-treatment and normal control cases

Enet		$L_1 + Fc.net$ *(gene)*		*HLR* ($L_{1/2} + L_2$)		$L_{1/2}$ + *central node Fc.net*	
cg1100973 (MARCO)	cg0237448 (PRF1)	**cg2079283 (PTPRCAP)**	cg15616083 (PAGE2)	cg02505409 (ANGPTL4)	cg21493583 (CRIPT)	**cg11804789 (CST7)**	cg06409153 (ABCA5)
cg0498897 (MPO)	cg2007009 (S100A8)	**cg04988978 (MPO)**	cg27303882 (MYL4)	cg06521852 (HRIHFB2122)	cg00201234 (FBLN2)	cg24505527 (NKIRAS2)	**cg05923103 (RNF11)**
cg2079283 (PTPRCAP)	cg2706761 (CYP4F3)	cg0996492 (KCNE1)	cg05294455 (ADORA1)	cg08694544 (RTBDN)	cg09638834 (RAET1L)	**cg15853125 (TIAM1)**	cg09497789 (SPAG17)
cg0996492 (KCNE1)	cg0435376 (MS4A6A)	cg11009736 (MARCO)	cg13626881 (ADORA1)	**cg15853125 (TIAM1)**	cg14861570 (MMD)	cg08694544 (RTBDN)	cg13626881 (ADORA1)
cg0652185 (HRIHFB2122)	cg0224062 (PLCB2)	cg14360917 (SP2)	**cg11412582 (HOXB5)**	cg21608192 (XYLT1)	cg09964921 (KCNE1)	cg07607462 (UBR1)	**cg11412582 (HOXB5)**
cg0013453 (UBASH3A)	cg0619637 (TREM1)	cg03801286 (KCNE1)	cg01405107 (IGLL1)	cg09497789 (SPAG17)	**cg04988978 (MPO)**	**cg20792833 (PTPRCAP)**	cg15736165 (BNC1)
cg1436091 (SP2)	cg21126943 (CEACAM6)	cg06521852 (HRIHFB2122)	cg10494770 (SNRPN)	**cg20792833 (PTPRCAP)**	cg14319409 (GLRA1)	**cg14027234 (CD248)**	**cg05105069 (TCEAL7)**
cg2193281 (CSTA)	cg0020123 (FBLN2)	cg21517055 (MGC11271)	cg24993443 (BRDG1)	cg06409153 (ABCA5)	cg26838900 (LRRC15)	cg00201234 (FBLN2)	cg07376232 (AMICA1)
cg0097486 (FCGR3B)	cg2746119 (FXYD1)	cg00201234 (FBLN2)	cg04398282 (ABCA5)	cg13626881 (ADORA1)	cg23490074 (C19orf2)	**cg04988978 (MPO)**	cg21493583 (CRIPT)
cg2151705 (MGC11271)	cg0529445 (MYL4)	cg00134539 (KCNQ2)	**cg14027234 (CD248)**	cg2193281 (CSTA)	cg17231524 (MGC39606)	cg06183267 (AFF3)	**cg03856723 (PRKACA)**

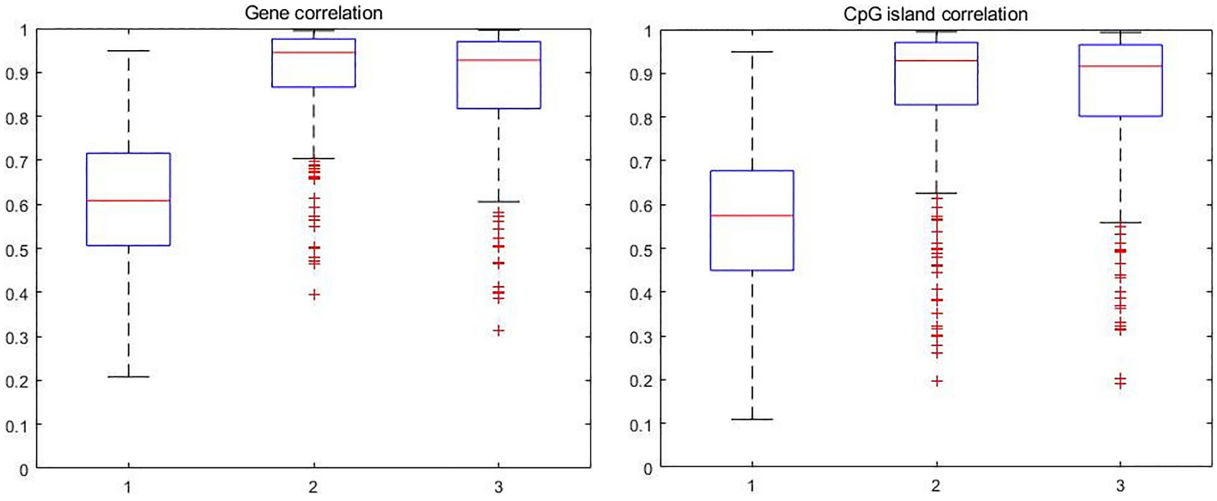

Fig. 4. The boxplot of correlation between CpG sites.

3 CpG sites in the data; samples with error were removed. Lastly, there were 156 controls case samples (Healthy sample), 120 pre-treatment case samples and 122 post-treatment case samples. For these three cases, we calculated the maximum correlation of CpG sites in each group (gene and CpG island).

Figure 3a–c shows the histogram of maximum sample correlation between CpG sites within genes in control, pre-treatment and post-treatment case where Fig. 3d–f shows the histogram of maximum sample correlation between CpG sites within CpG islands in control, pre-treatment and post-treatment cases. Figure 4 shows the boxplot of maximum sample correlation between CpG sites in gene or CpG islands. Based on Figs 3 and 4, the results show that most CpG sites within the same group have high correlation in pre-treatment case samples and post-treatment case samples whereas the control case samples only

Table 4
The top 20 CpG sites and the corresponding genes selected from the comparison between post-treatment and normal control cases

Enet		L_1 + *Fc.net (gene)*		*HLR* ($L_{1/2} + L_2$)		$L_{1/2}$ + *central node Fc.net*	
cg23580000 (ADCY7)	cg09626634 (EBI2)	cg06653796 (LIME1)	cg12243271 (CFI)	cg17682828 (FXYD7)	cg25554036 (WFS1)	cg11093356 (DDX19A)	cg04836428 (DTNA)
cg06653796 (LIME1)	cg22988566 (WFDC10B)	cg10986043 (TCAP)	cg10467098 (Bles03)	cg02713563 (TRAPPC6A)	cg23125689 (CD81)	**cg12711814 (ENO1)**	**cg10777851 (CD200)**
cg10986043 (TCAP)	cg24335895 (COX7A1)	cg23580000 (ADCY7)	cg19573166 (SLC22A17)	cg06653796 (LIME1)	**cg04232649 (CCNG1)**	cg12906740 (NUDT15)	cg00636639 (MRRF)
cg13379236 (EGF)	cg19573166 (SLC22A17)	cg13379236 (EGF)	cg15096140 (MYO1B)	cg15489301 (AKR1B10)	cg25410053 (ZIC3)	**cg14838256 (SRD5A2L)**	cg17133388 (C3orf28)
cg03547797 (GAS2)	cg15096140 (MYO1B)	cg03547797 (GAS2)	cg05767404 (C1orf150)	cg11093356 (DDX19A)	cg24643262 (BMX)	cg23002907 (RBMS2)	cg14275779 (PLEKHH3)
cg05135288 (RHOT2)	cg13745870 (SPATA12)	cg05135288 (RHOT2)	cg05004940 (C20orf195)	cg03547797 (GAS2)	cg26200585 (PRX)	cg02964389 (PSMD9)	cg07389922 (C17orf81)
cg20357806 (PPBP)	cg00134539 (UBASH3A)	cg12006284 (WT1)	cg23506842 (PTPN7)	cg20630655 (RNUT1)	cg14132995 (SLC35A2)	**cg23917399 (TNFAIP8)**	cg19514928 (TMEM56)
cg12006284 (WT1)	cg16853982 (ACTN2)	cg20357806 (PPBP)	cg23917399 (SPATA12)	cg10986043 (TCAP)	cg13056210 (MXRA5)	cg09119665 (PNMA1)	cg05798972 (PPARBP)
cg21640749 (CD300LF)	cg10467098 (Bles03)	cg24335895 (COX7A1)	cg09626634 (EBI2)	cg02497758 (MAFB)	cg04499381 (CXorf9)	cg17682828 (FXYD7)	cg00096922 (DLX5)
cg12243271 (CFI)	cg13247990 (MLCK)	cg21640749 (CD300LF)	**cg23917399 (TNFAIP8)**	cg25919221 (CA6)	cg13435792 (C12orf46)	cg09816912 (MARCKS)	cg04232649 (**CCNG1**)

show a significant correlation.

Table 2 shows the AUC for each method from real data analysis. In real data, the enhanced $L_{1/2}$ model also achieved higher accuracy rates. Tables 3 and 4 show the top 20 selected CpG sites for all methods. We further validated the chosen genes from the GeneCards Database (http://www.genecards.org). In Table 3, when comparing pre-treatment cases with controls, the algorithm $L_{1/2}+$ central node Fc.net (gene and CpG island) found various genes (TIAM1 [21,22], CST7 [25], TCEAL7 [23,24] and RNF11 [26]) that were not found by L_1+ *Fc.net (gene)* and Enet. Likewise, *HLR* ($L1/2 + L2$) was unable to find these genes (CST7, TCEAL7 and RNF11) where these genes (TIAM1, CST7, TCEAL7 and RNF11) were found to be correlated with cancer in previous research. On one hand, all methods were able to find genes (MPO [27,28], PTPRCAP [29]); on the other hand, network penalty methods were able to find genes (CD248 [31] and HOXB5 [32]). In Table 4, our algorithm $L_{1/2}+$ central node Fc.net (gene and CpG island) also found genes (CD200 [30], SRD5A2L [34], ENO1 [33]) which have not been found in L_1+ *Fc.net(gene)* and *Enet*. Additionally, gene CD200 and gene SRD5A2L also proved to be related to cancer. The L_1+ Fc.net and $L_{1/2}+$ central node Fc.net (gene and CpG island) algorithm, which has gene and island network information, also found gene (TNFAIP8 [35]) which was not found by *Enet/HLR*.

4. Conclusion

In biological molecular research, the analysis of DNA methylation may be a new practice for cancer research. In this paper, we used the enhanced $L_{1/2}$ penalized logistic regression model to extract divergently methylated CpG sites between healthy controls and ovarian cancer cases. We constructed the central node fully connected network which combines with genome information and CpG island information. We have advanced the corresponding coordinate descent algorithm suited for real DNA methylation data. This method not only has the $L_{1/2}$ penalty sparser than L_1, it also has more CpG sites relationship information. In real data, we used ovarian cancer samples with over 20,000 CpG sites. Even though the quantity of the

selected CpG sites was less than previous methods, more corresponding CpG sites within genes selected were potentially associated with cancers. Therefore, by comparing to traditional methods, our method clearly achieved a higher predictive accuracy. Therefore, the proposed method offers an advanced tool to researchers in DNA methylation and can be a powerful tool for recognizing pathogenic CpG sites.

Acknowledgments

The Macau Science and Technology Develop Funds (grant no. 003/2016 /AFJ) of Macau SAR of China supported this work.

Conflict of interest

None to report.

References

[1] Schöbeler D. Function and information content of DNA methylation. Nature. 2015; 517(7534): 321.
[2] Irizarry RA, Ladd-Acosta C, Wen B. et al. The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific CpG island shores. Nature Genetics. 2009; 41(2): 178-186.
[3] Baubec T, Colombo DF, Wirbelauer C. et al. Genomic profiling ofDNAmethyltransferases reveals a role for DNMT3B in genic methylation. Nature. 2015; 520(7546): 243.
[4] Pidsley R, Zotenko E, Peters TJ. et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biology. 2016; 17(1): 208.
[5] Bibikova M, Lin Z, Zhou L. et al. High-throughput DNA methylation profiling using universal bead arrays. Genome Research. 2006; 16(3): 383-393.
[6] Houseman EA, Christensen BC, Yeh RF. et al. Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. BMC Bioinformatics. 2008; 9(1): 365.
[7] Kuan PF, Wang S, Zhou X. et al. A statistical framework for Illumina DNA methylation arrays. Bioinformatics. 2010; 26(22): 2849-2855.
[8] Siegmund KD, Laird PW, Laird-Offringa IA. A comparison of cluster analysis methods using DNA methylation data. Bioinformatics. 2004; 20(12): 1896-1904.
[9] Wang S. Method to detect differentially methylated loci with case-control designs using Illumina arrays. Genetic Epidemiology. 2011; 35(7): 686-694.
[10] Friedman J, Hastie T, Höfling H. et al. Pathwise coordinate optimization. The Annals of Applied Statistics. 2007; 1(2): 302-332.
[11] Sun H, Wang S. Penalized logistic regression for high-dimensional DNA methylation data with case-control studies. Bioinformatics. 2012; 28(10): 1368-1375.
[12] Xu Z, Zhang H, Wang Y, Chang X, Liang Y. L1/2 regularization. Science China Information Sciences. 2010; 53(6): 1159-1169.
[13] Xu Z, Chang X, Xu F. et al. L1/2 regularization: a thresholding representation theory and a fast solver. IEEE Transactions on Neural Networks & Learning Systems. 2012; 23(7): 1013-27.
[14] Zeng J, Lin S, Wang Y. et al. L1/2 Regularization: Convergence of Iterative Half Thresholding Algorithm. IEEE Transactions on Signal Processing. 2013; 62(9): 2317-2329.
[15] Li F, Zhang NR. Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. Journal of the American Statistical Association. 2010; 105(491): 1202-1214.
[16] Tibshirani R. The lasso method for variable selection in the Cox model. Statistics in Medicine. 1997; 16(4): 385-395.
[17] Teschendorff AE, Menon U, Gentry-Maharaj A. et al. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Research. 2010; 20(4): 440-446.
[18] Bibikova M, Lin Z, Zhou L. et al. High-throughput DNA methylation profiling using universal bead arrays. Genome Research. 2006; 16(3): 383-393.

[19] Houseman EA, Christensen BC, Yeh RF. et al. Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. BMC Bioinformatics. 2008; 9(1): 365.
[20] Razin A, Cedar H. DNA methylation and gene expression. Microbiological Reviews. 1991; 55(3): 451-458.
[21] Liu L, Zhao L, Zhang Y, Zhang Q, Ding Y. Proteomic analysis of Tiam1-mediated metastasis in colorectal cancer. Cell Biology International. 2007; 31(8): 805-814.
[22] Liu L, Xu AG, Zhang QL, Zhang YF, Ding YQ. Effect of Tiam1 overexpression on proliferation and metastatic potential of human colorectal cancer. Zhonghua Bing Li Xue Za Zhi = Chinese Journal of Pathology. 2007; 36(6): 390-393.
[23] Peedicayil A, Vierkant RA, Shridhar V, Schildkraut JM, Armasu S, Hartmann LC. et al. Polymorphisms in TCEAL7 and risk of epithelial ovarian cancer. Gynecologic Oncology. 2009; 114(2): 260-264.
[24] Chien J, Staub J, Avula R, Zhang H, Liu W, Hartmann LC. et al. Epigenetic silencing of TCEAL7 (Bex4) in ovarian cancer. Oncogene. 2005; 24(32): 5089.
[25] Werle B, Schanzenbächer U, Lah TT, Ebert E, Jülke B, Ebert W. et al. Cystatins in non-small cell lung cancer: tissue levels, localization and relation to prognosis. Oncology Reports. 2006; 16(4): 647-655.
[26] Subramaniam V, Li H, Wong M, Kitching R, Attisano L, Wrana J. et al. The RING-H2 protein RNF11 is overexpressed in breast cancer and is a target of Smurf2 E3 ligase. British Journal of Cancer. 2003; 89(8): 1538.
[27] He C, Tamimi RM, Hankinson SE, Hunter DJ, Han J. A prospective study of genetic polymorphism in MPO, antioxidant status, and breast cancer risk. Breast Cancer Research and Treatment. 2009; 113(3): 585-594.
[28] Yang J, Ambrosone CB, Hong CC, Ahn J, Rodriguez C, Thun MJ, Calle EE. Relationships between polymorphisms in NOS3 and MPO genes, cigarette smoking and risk of post-menopausal breast cancer. Carcinogenesis. 2007; 28(6): 1247-1253.
[29] Ju H, Lim B, Kim M, Kim YS, Kim WH, Ihm C. et al. A regulatory polymorphism at position-309 in PTPRCAP is associated with susceptibility to diffuse-type gastric cancer and gene expression. Neoplasia. 2009; 11(12): 1340-1347.
[30] Moreaux J, Veyrune JL, Reme T, De Vos J, Klein B. CD200: a putative therapeutic target in cancer. Biochemical and Biophysical Research Communications. 2008; 366(1): 117-122.
[31] Simonavicius N, Robertson D, Bax DA, Jones C, Huijbers IJ, Isacke CM. Endosialin (CD248) is a marker of tumor-associated pericytes in high-grade glioma. Modern Pathology. 2008; 21(3): 308.
[32] Lee JY, Hur H, Yun HJ, Kim Y, Yang S, Kim SI, Kim MH. HOXB5 promotes the proliferation and invasion of breast cancer cells. International Journal of Biological Sciences. 2015; 11(6): 701.
[33] Qiao H, Wang YF, Yuan WZ, Zhu BD, Jiang L, Guan QL. Silencing of ENO1 by shRNA Inhibits the Proliferation of Gastric Cancer Cells. Technology in Cancer Research & Treatment. 2018; 17: 1533033818784411.
[34] Uemura M, Tamura K, Chung S, Honma S, Okuyama A, Nakamura Y, Nakagawa H. Novel 5α-steroid reductase (SRD5A3, type-3) is overexpressed in hormone-refractory prostate cancer. Cancer Science. 2008; 99(1): 81-8.
[35] Xing Y, Liu Y, Liu T, Meng Q, Lu H, Liu W. et al. TNFAIP8 promotes the proliferation and cisplatin chemoresistance of non-small cell lung cancer through MDM2/p53 pathway. Cell Communication and Signaling. 2018; 16(1): 43.