On the Identification of Correlated Differential Features for Supervised Classification of High-Dimensional Data

Shu Kay Ng and Geoffrey J. McLachlan

Abstract Many real problems in supervised classification involve high-dimensional feature data measured for individuals of known origin from two or more classes. When the dimension of the feature vector is very large relative to the number of individuals, it presents formidable challenges to construct a discriminant rule (classifier) for assigning an unclassified individual to one of the known classes. One way to handle this high-dimensional problem is to identify highly relevant differential features for constructing a classifier. Here a new approach is considered, where a mixture model with random effects is used firstly to partition the features into clusters and then the relevance of each feature variable for differentiating the classes is formally tested and ranked using cluster-specific contrasts of mixed effects. Finally, a non-parametric clustering approach is adopted to identify networks of differential features that are highly correlated. The method is illustrated using a publicly available data set in cancer research for the discovery of correlated biomarkers relevant to the cancer diagnosis and prognosis.

1 Introduction

In supervised classification, the data are classified with respect to g known classes and the intent is to construct a discriminant rule or classifier on the basis of these classified data for assigning an unclassified individual to one of the g classes on the basis of its feature vector. Many real problems in supervised classification, however, involve high-dimensional feature vectors. While there is a vast literature on dimensional reduction and/or feature selection in supervised classification [4, 8, 13],

S.K. Ng (✉)
School of Medicine and Menzies Health Institute Queensland, Griffith University, Nathan, QLD 4111, Australia
e-mail: s.ng@griffith.edu.au

G.J. McLachlan
Department of Mathematics, University of Queensland, St Lucia, QLD 4072, Australia
e-mail: g.mclachlan@uq.edu.au

F. Palumbo et al. (eds.), *Data Science*, Studies in Classification, Data Analysis, and Knowledge Organization, DOI 10.1007/978-3-319-55723-6_4

some of the methods may become inapplicable or unreliable when the dimension of the feature vector is very large relative to the number of individuals [2, 10, 15, 24]. An example of such an application is the analysis of gene-expression data, where expression levels of genes (features) are available from patients in g known classes of distinct disease stages or outcomes and the aim is to identify a small subset of "marker" genes that characterize the different classes and construct a discriminant rule to predict the class of origin of an unclassified patient [11, 17]. One way to handle this high-dimensional problem is to identify genes that are differentially expressed among the g classes of tissue samples. In this context, multiple hypothesis test-based approaches [27–29] have been proposed to assess statistical significance of differential expression for each gene separately, with control for the false discovery rate (FDR) which is defined as the expected proportion of false positives among the genes declared to be differentially expressed [1]. Clustering-based approaches have also been considered, but these methods either work on gene-specific summary statistics [14, 23] or reduced forms of gene-expression data [6]. Alternatively, clustering methods that can handle full gene-expression data rely on the assumption that pure clusters of null (non-differentially expressed) genes and differentially expressed genes exist [12, 26]; see also [25]. More recently, a mixture model-based approach with random-effects terms was proposed to draw inference on differences between classes using full gene-expression data [22]. This method does not rely on the clusters being pure as to whether all cluster members are differentially expressed or null genes. In this paper, we propose a new three-step method that extends this mixture model-based approach in order to identify networks of correlated differential features (genes) for supervised classification of high-dimensional data.

The rest of the paper is organized as follows. In Sect. 2, we describe the mixture model with random-effects terms [20] that is adopted in the first step to cluster the genes using full gene-expression data. We also present the second step, where the relevance of each feature variable for differentiating the classes is formally tested and ranked on the basis of cluster-specific contrasts of mixed effects. In Sect. 3, we describe the final third step in which a non-parametric clustering approach is used to further explore the group structures of selected highly ranked differential features for each cluster identified in the first step. Section 4 presents the application of the proposed method to a publicly available gene-expression data set in cancer research for the discovery of correlated biomarkers relevant to the cancer prognosis. Discussion is given in Sect. 5.

2 Mixture Model with Random-Effects Terms

With supervised classification, it is supposed that an individual belongs to one of g classes, denoted by $C_1, \ldots, C_g$, and that there is a vector of p feature variables measured on each individual. Based on the observed feature vectors, represented by an $n \times p$ matrix, the intent is to construct a discriminant rule for

allocating an unclassified individual to one of the g classes [15]. For applications in the context of supervised classification with gene-expression data, the number of individual tissue samples n is very small relative to the number of genes p. To handle this high-dimensional problem, it is proposed to adopt a mixture model with random-effects terms to firstly cluster the p genes and then identify those genes that are highly differentiated between the g classes of tissue samples.

Let $\boldsymbol{y}_j = (y_{1j}, \ldots, y_{nj})^T$ contain the measurements on the jth gene ($j = 1, \ldots, p$), where the superscript T denotes vector transpose and p is much greater than n. It is assumed that $\boldsymbol{y}_j$ has a h-component mixture distribution with probability π_i of belonging to the ith cluster ($i = 1, \ldots, h$), where the π_i sum to one. We let the h-dimensional vector $\boldsymbol{z}_j$ denote the cluster membership of $\boldsymbol{y}_j$, where $z_{ij} = (\boldsymbol{z}_j)_i = 1$ if $\boldsymbol{y}_j$ belongs to the ith cluster and zero otherwise ($i = 1, \ldots, h$). A mixture model with random-effects terms [20] is required because it is anticipated that repeated measurements of gene expression for a tissue sample and expression levels for a gene are both correlated; see also [19]. Specific random effects are thus considered in the mixture model to capture individual gene effects and the correlation between gene-expression levels among the tissue classes [22]. Conditional on its membership of the ith cluster, the distribution of $\boldsymbol{y}_j$ is specified by the linear mixed model

$$\boldsymbol{y}_j = \boldsymbol{X}\boldsymbol{\eta}_i + \boldsymbol{U}\boldsymbol{b}_{ij} + \boldsymbol{V}\boldsymbol{c}_i + \boldsymbol{\varepsilon}_{ij}, \tag{1}$$

where $\boldsymbol{X}, \boldsymbol{U}$, and $\boldsymbol{V}$ denote the known design matrices corresponding to the fixed effects terms $\boldsymbol{\eta}_i$ and to the random-effects terms $\boldsymbol{b}_{ij}$ and $\boldsymbol{c}_i$ ($i = 1, \ldots, h; j = 1, \ldots, p$), respectively. The vector $\boldsymbol{b}_{ij} = (b_{1ij}, \ldots, b_{gij})^T$ contains the unobservable gene-specific random effects for each of the g tissue classes, and $\boldsymbol{c}_i = (c_{1i}, \ldots, c_{ni})^T$ contains the random effects common to all genes from the ith cluster. The measurement error vector $\boldsymbol{\epsilon}_{ij}$ is taken to be multivariate normal $N_n(\boldsymbol{0}, \boldsymbol{A}_i)$, where $\boldsymbol{A}_i$ is a diagonal matrix. The vectors $\boldsymbol{b}_{ij}$ and $\boldsymbol{c}_i$ of random-effects terms are taken to be multivariate normal $N_g(\boldsymbol{0}, \boldsymbol{B}_i)$ and $N_n(\boldsymbol{0}, \boldsymbol{C}_i)$, respectively, where the variance component $\boldsymbol{C}_i$ is assumed to be diagonal and $\boldsymbol{B}_i$ is a non-diagonal $g \times g$ matrix, where the correlation between gene-specific random effects b_{lij} ($l = 1, \ldots, g$) is modelled via the off-diagonal elements in $\boldsymbol{B}_i$; see, for example, [22]. The assignment of the p genes into h clusters is implemented using the estimated conditional posterior probabilities of cluster membership given $\boldsymbol{y}_j$ and $\hat{\mathbf{c}}_l (j = 1, \ldots, p; l = 1, \ldots, g)$:

$$\tau_i(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}}, \hat{\mathbf{c}}) = \mathrm{pr}(Z_{ij} = 1 | \boldsymbol{y}_j, \hat{\mathbf{c}}) = \frac{\hat{\boldsymbol{\pi}}_i f(\boldsymbol{y}_j | z_{ij} = 1; \hat{\boldsymbol{\psi}}_i, \hat{\mathbf{c}}_i)}{\sum_{m=1}^{h} \hat{\boldsymbol{\pi}}_m f(\boldsymbol{y}_j | z_{mj} = 1; \hat{\boldsymbol{\psi}}_m, \hat{\mathbf{c}}_m)}, \tag{2}$$

where $\boldsymbol{\psi}_i$ is the parameter vector for the ith component density containing the unknown parameters $\boldsymbol{\eta}_i$ and distinct elements in $\boldsymbol{A}_i$, $\boldsymbol{B}_i$, and $\boldsymbol{C}_i (i = 1, \ldots, h)$, and

$$\log f(\boldsymbol{y}_j | z_{ij} = 1; \hat{\boldsymbol{\psi}}_i, \hat{\mathbf{c}}_i) = -\frac{1}{2} \Big\{ \log |\hat{\mathbf{D}}_i| + (\boldsymbol{y}_j - \boldsymbol{X}\hat{\boldsymbol{\eta}}_i - \boldsymbol{V}\hat{\mathbf{c}}_i)^T \hat{\mathbf{D}}_i^{-1} (\boldsymbol{y}_j - \boldsymbol{X}\hat{\boldsymbol{\eta}}_i - \boldsymbol{V}\hat{\mathbf{c}}_i) \Big\}$$

is the log density of $\boldsymbol{y}_j$ conditioned on $\hat{\mathbf{c}}_i$ and the membership of the ith cluster, apart from an additive constant, and where $\hat{\mathbf{D}}_i = \hat{\mathbf{A}}_i + \boldsymbol{U}\hat{\mathbf{B}}_i\boldsymbol{U}^T$; see [20].

To quantify the relevance of each gene for differentiating the g classes, we consider an individual observation-specific contrast in the estimates of the fixed and random effects weighted by the estimated posterior probabilities (2) of cluster membership:

$$W_j = \sum_{i=1}^{h} \tau_i(\boldsymbol{y}_j; \hat{\boldsymbol{\Psi}}, \hat{\mathbf{c}})\, \hat{S}_{ij} \qquad (j = 1, \ldots, p), \tag{3}$$

where

$$\hat{S}_{ij} = \boldsymbol{d}_j^T (\hat{\boldsymbol{\eta}}_i^T, \hat{\mathbf{b}}_{G_i}^T, \hat{\mathbf{c}}_i^T)^T \Big/ \sqrt{\boldsymbol{d}_j^T \hat{\boldsymbol{\Omega}}_i \boldsymbol{d}_j} \tag{4}$$

is the cluster-specific normalized contrast with the BLUP estimator of the mixed effects, and where $\boldsymbol{d}_j$ is a vector whose elements sum to zero, $\boldsymbol{b}_{G_i} = (\boldsymbol{b}_{i_1}^T, \ldots, \boldsymbol{b}_{i_{p_i}}^T)^T$ contains the gene-specific random-effects terms for the p_i genes belonging to the ith cluster G_i $(i = 1, \ldots, h)$, and $\hat{\boldsymbol{\Omega}}_i$ is the covariance matrix of the BLUP estimator of the mixed effects, which can be partitioned conformally corresponding to $\boldsymbol{\eta}_i | \boldsymbol{b}_{G_i} | \boldsymbol{c}_i$, respectively, as described in [22].

Based on the weighted contrast W_j $(j = 1, \ldots, p)$ given in (3), the p genes can be ranked in the order of their relevance for differentiating the g classes (with respect to the defined form of d_j for the normalized contrast (4)). In the final step of the proposed method to be described in the next section, we intend to explore the group structure of top-ranked differentially expressed genes in each identified cluster $G_i (i = 1, \ldots, h)$, say, for those genes with contrast W_j more extreme than thresholds w_{0u} or w_{0d} for upregulated and downregulated genes, respectively. A guide to plausible values of w_{0u} and w_{0d} can be obtained using the percentile rank of $W_j (j = 1, \ldots, p)$, whereby the percentiles are taken to be the mixing proportions of the non-central portions of W_j fitted by a three-component mixture of t-distributions (these two components are considered as representing the distribution of W_j for upregulated and downregulated differentially expressed genes).

3 A Non-parametric Clustering Approach for Identification of Correlated Features

We consider the r_i top-ranked genes with W_j more extreme than either w_{0u} or w_{0d} in Cluster $G_i (i = 1, \ldots, h)$ and adopt a non-parametric method to cluster the r_i genes into networks of differentially expressed genes that are highly correlated. The method starts with the calculation of pairwise correlation coefficients for each pair of the r_i genes in $G_i (i = 1, \ldots, h)$. Significance of the pairwise correlation coefficients is then assessed with the use of a permutation method [21] to determine the null distribution of correlation coefficients. Precisely, the n class labels of tissue samples are randomly permuted separately for each gene. We pool the permutations for all $N_{r_i} = r_i(r_1 - 1)/2$ pairs of genes to determine the null distribution of correlation coefficients. In this paper, we consider the use of $S = 100$ repetitions of permutations and estimate the P-value for each pair of genes by

$$P_l = \sum_{s=1}^{S} \frac{\#\{m : R_{0m}^{(s)} \geq R_l, m = 1, \ldots, N_{r_i}\}}{N_{r_i} S} \qquad (l = 1, \ldots, N_{r_i}), \tag{5}$$

where $R_{0m}^{(s)}$ is the null version of correlation coefficient for the mth pair of genes after the sth repetition of permutations ($m = 1, \ldots, N_{r_i}$; $s = 1, \ldots, S$). Let $P_{(1)} \leq \cdots \leq P_{(N_{r_i})}$ be the ordered observed P-values obtained from (5). The Benjamini–Hochberg procedure [1] is adopted to determine the cut-off $\hat{k}$, where

$$\hat{k} = \arg\max\{k : P_{(k)} \leq \alpha k / N_{r_i}\}, \tag{6}$$

with control of the FDR at level α. Pairwise correlation coefficients corresponding to P-values $P_{(1)} \leq \cdots \leq P_{(\hat{k})}$ are identified to be significant. Significance of the pairwise correlation coefficients is represented by an $r_i \times r_i$ symmetric binary matrix M with elements of one or zero indicating that the corresponding correlation coefficients are significance or not. Finally, we search in M to identify networks of differentially expressed genes in which all members in a group significantly correlate with one another [21]. This non-parametric clustering approach obtains overlapping groups (networks) of correlated differentially expressed genes.

4 Real Example

We consider the colorectal cancer gene-expression data set [5], which comprised expression values of 15,552 genes for plasma samples from 12 colorectal cancer patients and 8 healthy donors. The original study aims to validate the power of four randomly selected markers (from a list of 40 genes differentially upregulated

in cancer patients) in enabling differentiation of the tumour from the healthy condition [5]. With the proposed three-step approach, we first fitted a mixture model with random-effects terms to the column-normalized gene-expression data set with h=3 to h=20 clusters, taking $\boldsymbol{X} = \boldsymbol{U}$ to be a 20×2 zero-one matrix (the first 12 rows are $(1, 0)$ and the next 8 rows are $(0, 1)$) and taking $\boldsymbol{V}$ to be $\boldsymbol{I}_{20}$. Based on the Bayesian information criterion (BIC) for model selection, we identified that there are 15 clusters of genes. The ML estimates of the unknown parameters are presented in Table 1. The ranking of differentially expressed genes is then implemented on the basis of the weighted estimates of a contrast in the mixed effects (3). For the case of g=2 classes of tissue samples (tumour versus healthy), we consider $\boldsymbol{d}_j$ of the form as

$$\boldsymbol{d}_j^T = (1 \ -1 \vdots 0 \ 0, \ \ldots, \ 0 \ 0, \ 1 \ -1, \ 0 \ 0, \ \ldots \vdots 0 \ \ldots \ 0), \tag{7}$$

where only one pair of $(1 \ -1)$ exists in the second partition corresponding to $\boldsymbol{b}_{G_i}$; see Eq. (4). We then fitted a three-component mixture of t-distributions [16] to W_j and obtained the mixing proportions of the components corresponding to the non-central portion of W_j, which are 11.5 and 7.2% for upregulated and downregulated genes in the tumour tissues, respectively. Thus we selected $w_{0u} = 1.661$ (the 88.5th percentile of W_j) and $w_{0d} = -2.236$ (the 7.2th percentile of $W_j (j = 1, \ldots, p)$). There are a total of 2907 differentially expressed genes with W_j more extreme than w_{0u} or w_{0d} ($W_j > 1.661$ or $W_j < -2.236$). Among them, 1581 genes have valid identifiers (1073 upregulated

Table 1 Estimates of the mixture model with random-effects terms for the colorectal cancer data set (15 clusters)

i	π_i	η_i (η_{1i}, η_{2i})	A_i $(\sigma_{1i}, \sigma_{2i})$	B_i $(\sigma_{b1i}, \sigma_{b2i}, \sigma_{b12i})$	C_i (σ_{ci})
1	0.024	−0.601, −0.591	0.643, 0.919	0.144, 0.085, 0.105	0.020
2	0.031	0.144, 0.239	0.954, 2.790	0.033, 0.095, 0.039	0.006
3	0.109	−0.232, −0.030	0.578, 0.511	0.024, 0.029, 0.023	0.011
4	0.035	0.032, 0.059	1.721, 0.336	0.054, 0.035, 0.013	0.001
5	0.114	0.070, 0.126	0.646, 0.235	0.055, 0.053, 0.031	0.004
6	0.036	0.217, 0.466	0.382, 0.567	0.035, 0.060, 0.037	0.004
7	0.092	−0.043, −0.299	0.265, 0.480	0.050, 0.037, 0.036	0.024
8	0.092	−0.004, −0.067	0.664, 1.304	0.028, 0.066, 0.030	0.000
9	0.026	−0.052, −0.535	1.622, 2.926	0.215, 0.156, 0.165	0.068
10	0.104	0.034, 0.094	1.487, 1.590	0.039, 0.088, 0.039	0.008
11	0.034	0.151, 0.037	2.770, 1.602	0.087, 0.119, 0.048	0.014
12	0.069	−0.418, −0.173	0.905, 0.715	0.022, 0.023, 0.018	0.034
13	0.130	0.211, 0.151	1.314, 0.859	0.036, 0.094, 0.036	0.002
14	0.034	0.454, −0.016	0.680, 0.691	0.019, 0.053, -0.002	0.020
15	0.070	0.012, 0.081	0.179, 0.196	0.049, 0.071, 0.048	0.004

Table 2 Descriptive statistics of W_j for the differentially expressed genes with valid gene identifiers and W_j more extreme than either w_{0u} or w_{0d} (15 clusters)

i	r_i	Mean (SD)	Median (IQR)	(Minimum, maximum)
1	1	−2.269 (n.a.)	−2.269 (n.a.)	N.a.
2	0	N.a.	N.a.	N.a.
3	173	−2.412 (0.127)	−2.378 (0.195)	(−2.782, −2.240)
4	0	N.a.	N.a.	N.a.
5	27	−0.124 (2.167)	1.679 (4.232)	(−2.486, 2.436)
6	44	−2.471 (0.194)	−2.458 (0.255)	(−3.052, −2.239)
7	714	2.391 (0.406)	2.429 (0.639)	(1.662, 3.635)
8	2	1.772 (0.005)	1.772 (n.a.)	(1.768, 1.776)
9	101	2.231 (0.415)	2.160 (0.567)	(1.669, 3.244)
10	1	1.775 (n.a.)	1.775 (n.a.)	N.a.
11	4	1.996 (0.142)	2.019 (0.262)	(1.803, 2.142)
12	264	−2.607 (0.161)	−2.725 (0.221)	(−2.940, −2.237)
13	10	1.856 (0.186)	1.696 (0.277)	(1.667, 2.185)
14	224	2.339 (0.468)	2.249 (0.719)	(1.667, 3.873)
15	16	−1.660 (1.706)	−2.362 (0.242)	(−2.772, 1.876)

Notation: *SD* standard deviation, *IQR* interquartile range, *n.a.* not appropriate

and 508 downregulated). Descriptive statistics of W_j for these 1581 differentially expressed genes are provided in Table 2. It can be seen that Clusters 7–11 and 13–14 contain upregulated differentially expressed genes, Clusters 1, 3, 6, and 12 contain downregulated differentially expressed genes, and Clusters 5 and 15 contain both upregulated and downregulated differentially expressed genes.

In the final step, we applied the non-parametric method to identify networks of correlated differentially expressed genes from the r_i genes in Cluster G_i. We set α to be between 0.1 and 0.00005 such that the expected number of false positives among the pairs of genes identified to be significantly correlated is smaller than one; see [21]. With the matrix M, networks of differentially expressed genes were displayed using UCINET6 for Windows [3]. Figure 1 presents the identified networks of upregulated differentially expressed genes in Clusters 7, 9, 13, and 14, where the nodal size of a gene is proportional to the degree of the node (the number of genes that are significantly correlated with the gene). Networks of downregulated differentially expressed genes (Clusters 3, 6, and 12) were provided in Fig. 2. Clusters 5 and 15 had networks of up- and down regulated differentially expressed genes (Fig. 3).

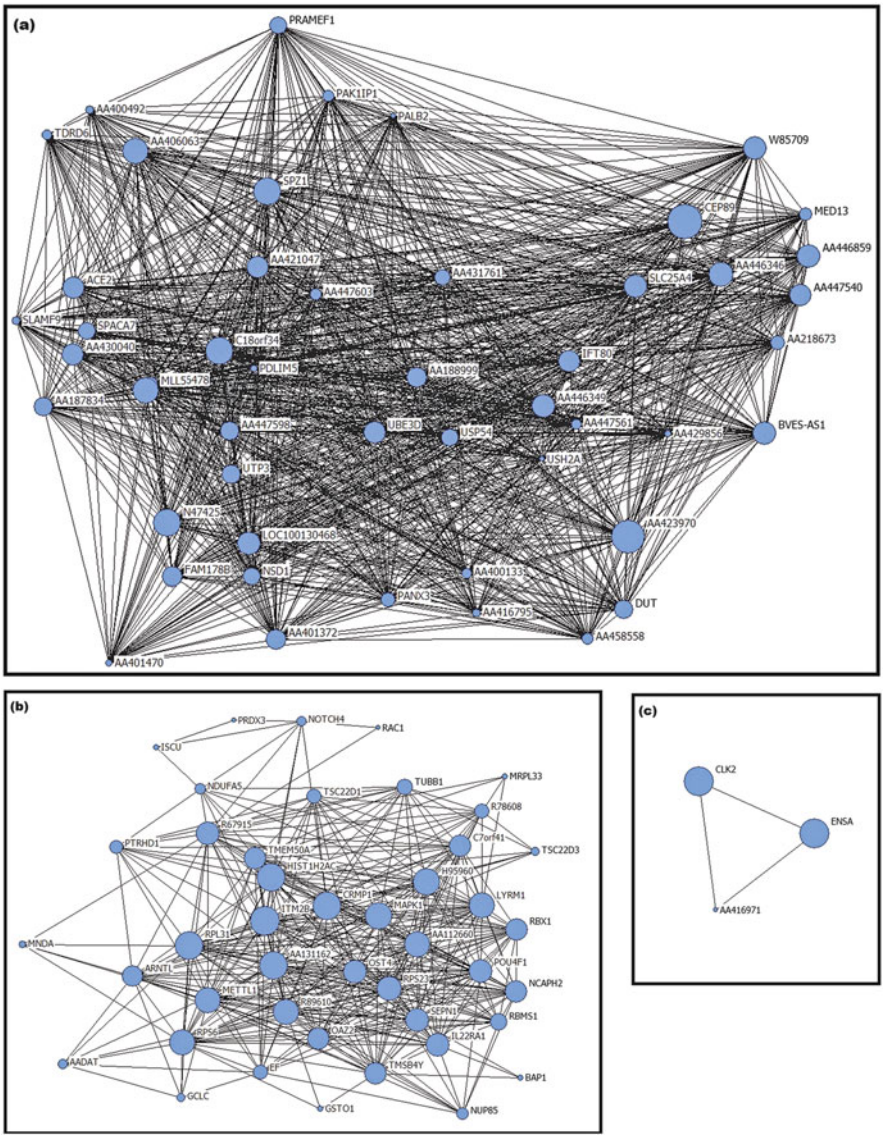

Fig. 1 Network of upregulated differentially expressed genes in (**a**) Cluster 7; (**b**) Cluster 9; (**c**) Cluster 13; and (**d**) Cluster 14. Nodal size is proportional to the degree (the number of genes that are significantly correlated with the gene). For Clusters 7, 9, and 13, only genes with the top 50 degrees were displayed

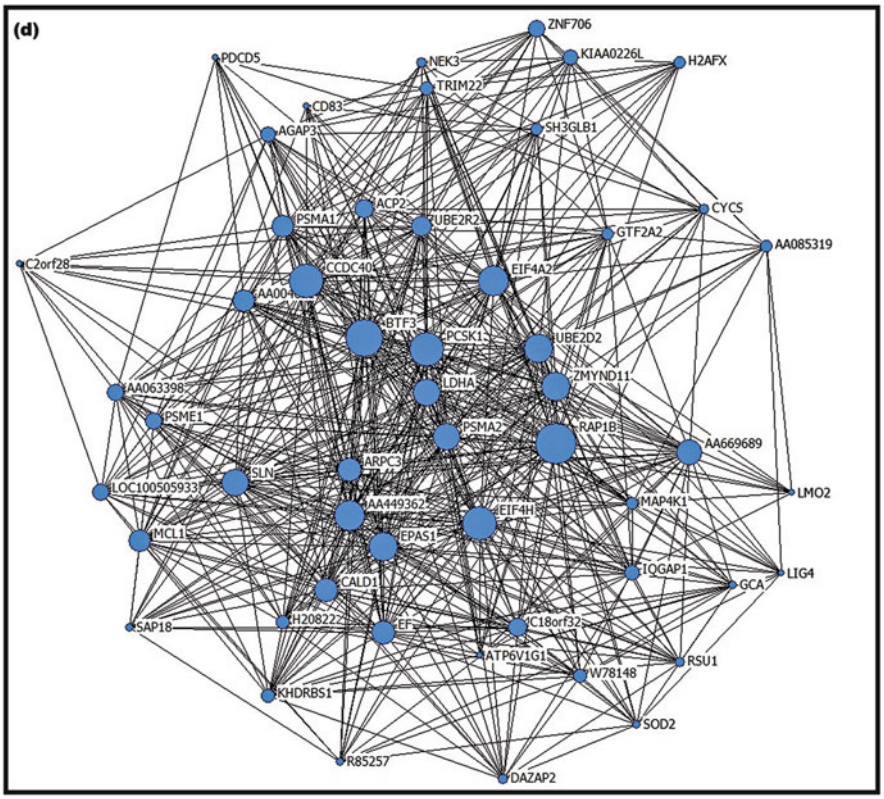

Fig. 1 (continued)

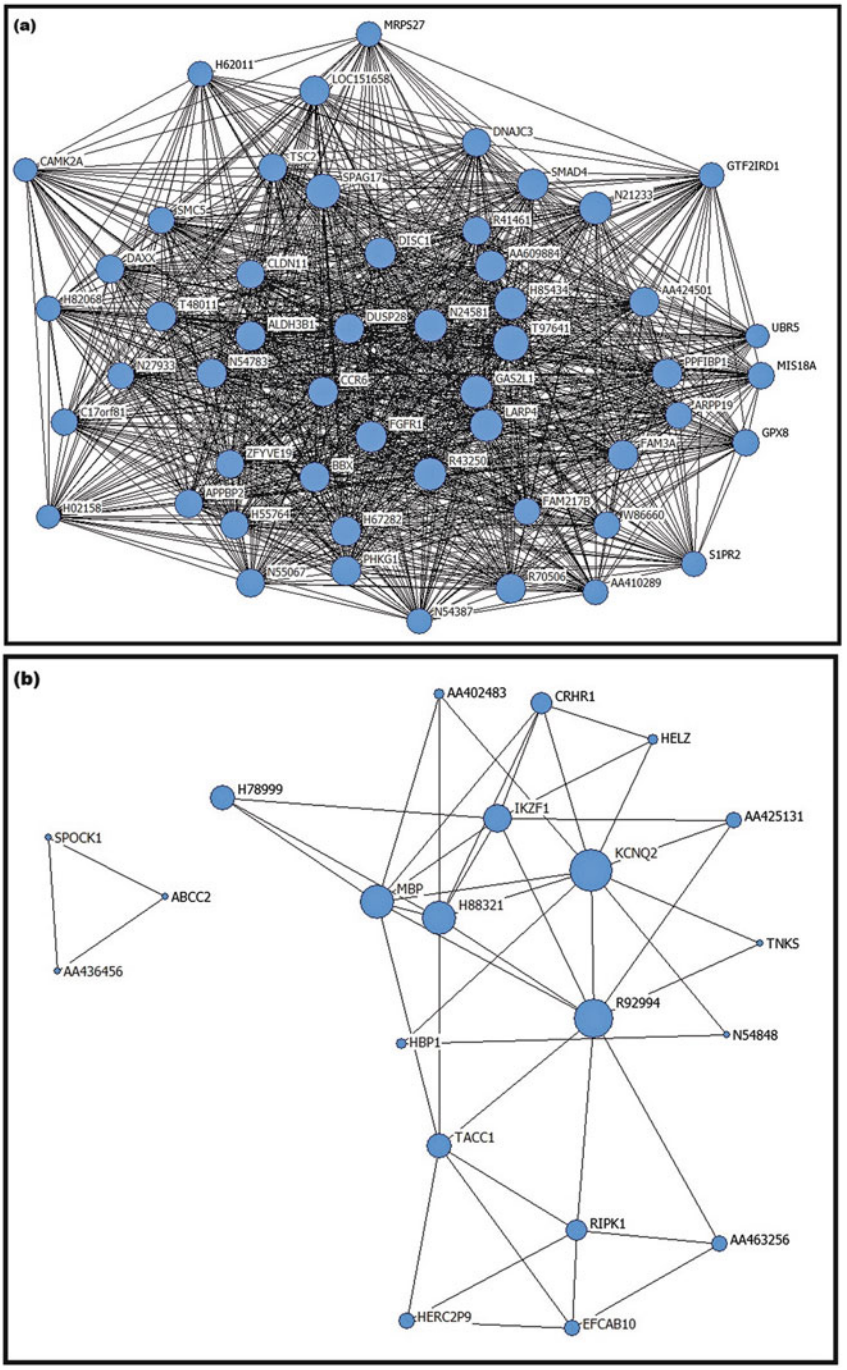

Fig. 2 Network of downregulated differentially expressed genes in (**a**) Cluster 3; (**b**) Cluster 6; and (**c**) Cluster 12. Nodal size is proportional to the degree (the number of genes that are significantly correlated with the gene). For Clusters 3 and 12, only genes with the top 50 degrees were displayed

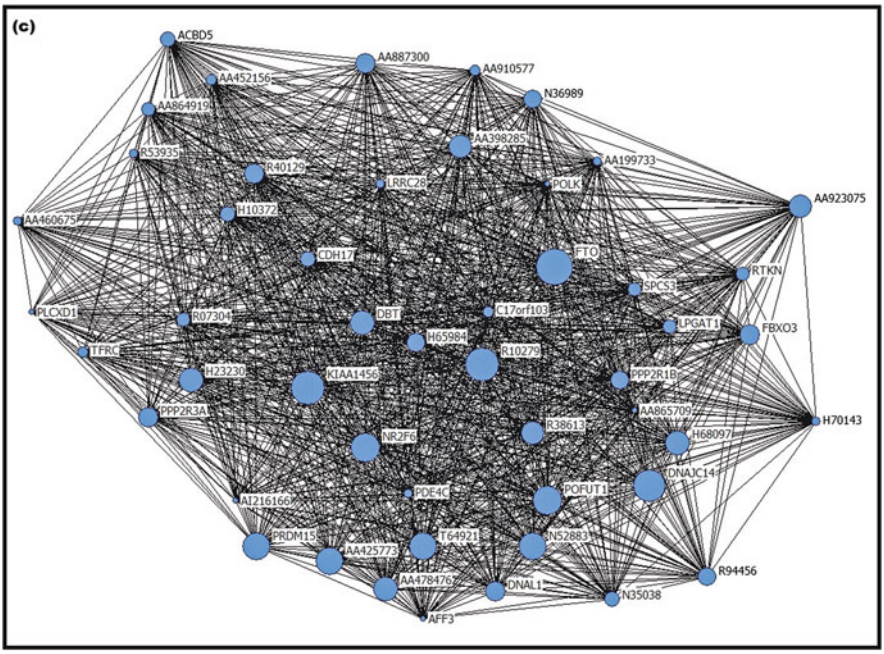

Fig. 2 (continued)

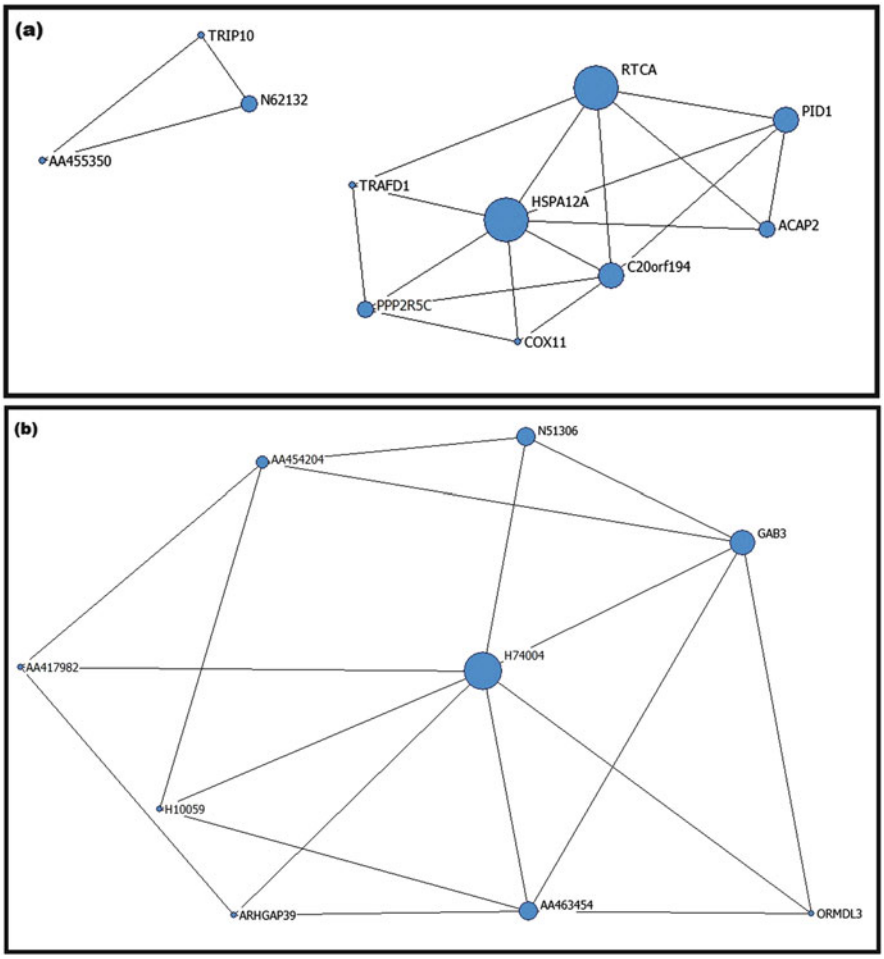

Fig. 3 Network of upregulated and downregulated differentially expressed genes in (**a**) Cluster 5 and (**b**) Cluster 15. Nodal size is proportional to the degree (the number of genes that are significantly correlated with the gene). From (**a**), it can be seen that two separate networks were identified for genes belonging to Cluster 5. One of them contains upregulated genes HSPA12A, RTCA, PID1, C20orf194, ACAP2, PPP2R5C, COX11, and TRAFD1. Another one contains downregulated genes N62132, AA455350, and TRIP10. For Cluster 15 (**b**), the network contains genes that are downregulated except ARHGAP39 (upregulated), which significantly correlated with downregulated genes {H74004, AA417982} and {H74004, AA463454}

A summary of the identified networks of correlated differentially expressed genes for each cluster is given in Table 3. Two isolated networks of differentially expressed genes were identified: {N62132, TRIP10, AA455350} downregulated genes network from Cluster 5 and {CLK2, ENSA, AA416971} upregulated genes network from Cluster 13. It is noted that four upregulated genes were considered

Table 3 A summary of networks of highly correlated differentially expressed genes

i	r_i	Highly correlated differentially expressed genes
1	1	NRGN
3	173[a]	T97641, SPAG17, N24581, LARP4, N21233, R43250, GAS2L1, H85434, DISC1, FGFR1
5	27[b]	(N62132, TRIP10, AA455350), HSPA12A, RTCA, PID1, C20orf194, ACAP2, PPP2R5C, COX11, TRAFD1
6	44[a]	KCNQ2, R92994, H88321, MBP, IKZF1, H78999, TACC1, RIPK1, CRHR1, AA463256, AA425131, EFCAB10, HERC2P9
7	714[a]	CEP89, AA423970, C18orf34, N47425, SPZ1, MLL5, AA406063, AA446346, AA446349, AA446859, W85709
8	2	SELT, PIP4K2A
9	101[a]	ITM2B, CRMP1, AA131162, HIST1H2AC, RPL31, MAPK1, H95960, R89610, LYRM1, AA112660, RPS6, METTL1
10	1	RGS2
11	4	MTG1, QK1, EF, SPARCL1
12	264[a]	FTO, KIAA1456, R10279, DNAJC14, POFUT1, NR2F6, PRDM15, T64921, N52883, AA425773
13	10[b]	(CLK2, ENSA, AA416971)
14	224[a]	RAP1B, BTF3, CCDC40, PCSK1, EIF4H, EIF4A2, AA449362, ZMYND11, EPAS1, UBE2D2
15	16	H74004, GAB3, AA463454, N51306, AA454204, ORMDL3, CPSF6, H10059, AA417982, ARHGAP39

[a]For large networks, only genes with the top ten degrees were listed
[b]Genes that form an isolated network are grouped within a bracket (in Clusters 5 and 13)

in the original study and three of them (EPAS1, UBE2D3, KIAA0101) were validated to be significantly increased in cancer compared to healthy donors [5]. Our clustering results confirmed the same findings; these three genes were identified as differentially expressed genes in Cluster 14 (with contrast W_j = 3.7, 3.3, and 2.0, respectively, and ranked the 2nd, 8th, and 156th among the 224 differentially expressed genes in Cluster 14). The original study could not validate the remaining upregulated gene DDX46. However, our method has sufficient power to identify DDX46 as a differentially expressed gene in Cluster 5, with $W_j = 2.4$ and ranked the 1st among the 14 upregulated differentially expressed genes in Cluster 5.

5 Discussion

We have presented a new approach to identify correlated differential features for supervised classification of high-dimensional data. The method adopts a mixture model with random-effects terms to cluster the feature variables and then ranks them in terms of their cluster-specific contrasts of mixed effects that quantify the evidence of differentiation between the known classes. The final step of the method adopts a

non-parametric clustering approach to identify networks of differential features that are highly correlated in each identified cluster.

The proposed method is illustrated using an application on the analysis of gene-expression cancer data. The identified differentially expressed genes and their correlation structures can have significant contribution in the discovery of novel biomarkers relevant to the cancer diagnosis and prognosis; see also [7, 9] for the benefit of using the covariance information among genes for feature selection. Moreover, these differentially expressed genes can be included in a model to construct a classifier with a smaller subset of marker genes, using methods such as mixtures of factor analysers [15, 16] or mixtures of multivariate generalized Bernoulli distributions [18]. This work will be pursued in future research.

Acknowledgements Part of this work has been presented in the Conference of the International Federation of Classification Societies, Bologna, July 2015. This work was supported by a grant from the Australian Research Council.

References

1. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B **57**, 259–300 (1995)
2. Bickel, P.J., Levina, E.: Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. Bernoulli **10**, 989–1010 (2004)
3. Borgatti, S.P., Everett, M.G., Freeman, L.C.: Ucinet for Windows: Software for Social Network Analysis. Analytic Technologies, Harvard, MA (2002). Available via http://www.analytictech.com/. Accessed 8 Dec 2015
4. Cai, T., Liu, W.: A direct estimation approach to sparse linear discriminant analysis. J. Am. Stat. Assoc. **106**, 1566–1577 (2011)
5. Collado, M., Garcia, V., Garcia, J.M., Alonso, I., Lombardia, L., et al.: Genomic profiling of circulating plasma RNA for the analysis of cancer. Clin. Chem. **53**, 1860–1863 (2007)
6. Dahl, D.B., Newton, M.A.: Multiple hypothesis testing by clustering treatment effects. J. Am. Stat. Assoc. **102**, 517–526 (2007)
7. Donoho, D., Jin, J.: Higher criticism for large-scale inference, especially for rare and weak effects. Stat. Sci. **30**, 1–25 (2015)
8. Fan, J., Lv, J.: A selective review of variable selection in high dimensional feature space. Stat. Sin. **20**, 101–148 (2010)
9. Fan, J., Feng, Y., Tong, X.: A road to classification in high dimensional space: the regularized optimal affine discriminant. J. R. Stat. Soc. B **74**, 745–771 (2012)
10. Hall, P., Pittelkow, Y., Ghosh, M.: Theoretic measures of relative performance of classifiers for high-dimensional data with small sample sizes. J. R. Stat. Soc. B **70**, 158–173 (2008)
11. Hall, P., Jin, J., Miller, H.: Feature selection when there are many influential features. Bernoulli **20**, 1647–1671 (2014)
12. He, Y., Pan, W., Lin, J.: Cluster analysis using multivariate normal mixture models to detect differential gene expression with microarray data. Comput. Stat. Data Anal. **51**, 641–658 (2006)
13. Kersten, J.: Simultaneous feature selection and Gaussian mixture model estimation for supervised classification problems. Pattern Recogn. **47**, 2582–2595 (2014)

14. Matsui, S., Noma, H.: Estimating effect sizes of differentially expressed genes for power and sample-size assessments in microarray experiments. Biometrics **67**, 1225–1235 (2011)
15. McLachlan, G.J.: Discriminant analysis. WIREs Comput. Stat. **4**, 421–431 (2012)
16. McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley, New York (2000)
17. McLachlan, G.J., Do, K.A., Ambroise, C.: Analyzing Microarray Gene Expression Data. Wiley, New York (2004)
18. Ng, S.K.: A two-way clustering framework to identify disparities in multimorbidity patterns of mental and physical health conditions among Australians. Stat. Med. **34**, 3444–3460 (2015)
19. Ng, S.K., McLachlan, G.J.: Mixture models for clustering multilevel growth trajectories. Comput. Stat. Data Anal. **71**, 43–51 (2014)
20. Ng, S.K., McLachlan, G.J., Wang, K., Ben-Tovim, L., Ng, S.-W.: A mixture model with random-effects components for clustering correlated gene-expression profiles. Bioinformatics **22**, 1745–1752 (2006)
21. Ng, S.K., Holden, L., Sun, J.: Identifying comorbidity patterns of health conditions via cluster analysis of pairwise concordance statistics. Stat. Med. **31**, 3393–3405 (2012)
22. Ng, S.K., McLachlan, G.J., Wang, K., Nagymanyoki, Z., Liu, S., Ng, S.-W.: Inference on differences between classes using cluster-specific contrasts of mixed effects. Biostatistics **16**, 98–112 (2015)
23. Pan, W., Lin, J., Le, C.T.: Model-based cluster analysis of microarray gene-expression data. Genome Biol. **3**, 0009.1–0009.8 (2002)
24. Pyne, S., Lee, S.X., Wang, K., Irish, J., Tamayo, P., et al.: Joint modeling and registration of cell populations in cohorts of high-dimensional flow cytometric data. PLoS One **9**, e100334 (2014)
25. Qi, Y., Sun, H., Sun, Q., Pan, L.: Ranking analysis for identifying differentially expressed genes. Genomics **97**, 326–329 (2011)
26. Qiu, W., He, W., Wang, X., Lazarus, R.: A marginal mixture model for selecting differentially expressed genes across two types of tissue samples. Int. J. Biostat. **4**, Article 20 (2008)
27. Smyth, G.: Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. **3**, Article 3 (2004)
28. Storey, J.D.: The optimal discovery procedure: a new approach to simultaneous significance testing. J. R. Stat. Soc. B **69**, 347–368 (2007)
29. Zhao, Y.: Posterior probability of discovery and expected rate of discovery for multiple hypothesis testing and high throughput assays. J. Am. Stat. Assoc. **106**, 984–996 (2011)