Innovative hybridisation of genetic algorithms and neural networks in detecting marker genes for leukaemia cancer

Dong Ling Tong[1], Keith Phalp[1], Amanda Schierz[1] and Robert Mintram[2]

[1] Bournemouth University, School of Design, Engineering and Computing, Poole House, Talbot Campus, Fern Barrow, Poole, Dorset, BH12 5BB, UK
[2] Personal contact

Abstract. The hybridisation of genetic algorithm (GAs) and artificial neural network (ANNs) are not new for microarray studies. However, these hybrid systems require data preprocessing and focus on classification accuracy. In this paper, a feature selection method based on the hybridisation of GAs and ANNs for oligonucleotide microarray data is proposed. The fitness values of the GA chromosomes are defined as the number of correctly labelled samples returned by the ANN. For gene validation, three supervised classifiers have been employed to evaluate the significance of selected genes based on a separate set of unknown sample data. Experimental results show that our method is able to extract informative genes without data preprocessing and this has reduced the gene variability errors in the selection process.

1 Introduction

The main challenge when working with microarray data is to identify computationally effective and biologically meaningful analysis models that extract the most informative and unbiased marker genes from a large pool of genes that are not involved in the array experiments. Numerous genetic algorithm (GA) and neural network (ANN) hybrid systems have been developed to emphasize effective classification [1–3, 8, 11]. Beiko and Charlebois [1] utilised the evolutionary ability of GAs to identify the best combinations of sequence indices and ANN architecture for DNA sequence classification. Meanwhile, Karzynski et al. [8] used GAs to optimise both the architecture and weight assignment of ANNs for multiclass recognition. In addition to optimising ANNs using GAs, recent studies tend to utilise the universal computational power of ANNs to compute GA fitness function for cancer classification. For instance, Bevilacqua et al. [2] and Lin et al. [11] applied the error rates returned by ANN to determine the fitness of GA chromosome in the classification of breast cancer metastasis recurrence and multiclass microarray datasets, respectively. Cho et al. [3] defined GA fitness based on ANN classification results on SRBCTs tumour data.

Instead of improving classification performance, our work is focused on feature selection for oligonucleotide microarray data using GAs and ANNs. For gene selection, the fitness values of GA chromosomes are defined as the number of correctly labelled samples returned by a feedforward ANN. To show the selection performance, the same experimental dataset as Golub et al. [7] is used. For gene validation, some commonly applied classifiers, such as multilayer perceptron (MLP), support vector machine (SVM) and nearest neighbour (KNN) have been employed using the WEKA suite of data mining software and the selection results are compared with previous experiments based on the same experimental dataset.

The rest of the paper is organised as follows. Section 2 describes proposed feature selection model. The experiment results are presented in Section 3. Finally in Section 4, the conclusion is drawn.

2 Methods

2.1 Feature Selection Model

The proposed model has three components: population initialisation, fitness computation and pattern evaluation.

Population initialisation A set of features from the experimental dataset is randomly chosen by the GA. These features form a finite feature space and are used as a basis for the fitness computation of each member of the population. To find the optimal population size, preliminary experiments were conducted on 4 variants of population sizes: 50, 100, 200 and 300. The population size of 300 was found to be the most optimal. Each chromosome is represented by 10 genes expressions which is encoded with a real number representation.

Fitness computation The fitness function is the number of correctly labelled samples returned by the ANN as shown in equation Eq. 1. For proposed method, chromosomes with higher fitness values are more likely to survive than those with least values.

$$\text{fitness} = \sum_{i=1}^{n} \sum_{k=1}^{c} s_{ik}, \tag{1}$$

$$s_{ik} = t_{ik} - \sqrt{(a_{ik} - c_{ik})^2} \quad \begin{cases} 1, & o_{ik} = t_{ik} \\ 0, & o_{ik} \neq t_{ik} \end{cases}, \tag{2}$$

$$C_k = \frac{1}{s_k} \sum_{s \epsilon k} a_{sk}, \tag{3}$$

where s_{ik}, t_{ik}, a_{ik}, o_{ik} and c_k represent sample data, target value of the sample, network activation output, actual output value and class centroid value, respectively. A feedforward ANN with the structure of 10-5-2 and network size of 67 including 5 and 2 bias nodes in the hidden and output layers, respectively, is constructed. The *tanh* activation function is employed as it is one of the commonly used nonlinear functions in the ANN.

Pattern evaluation Two sets of evaluation were performed using the GA: feature evaluation and network evaluation. For feature evaluation, two parent chromosomes are crossovered to produce a new offspring which is then mutated to create diversity from its parent. For network evaluation, two parent networks are crossovered to form a new set of network weights which then will be used to compute the fitness value of feature offspring. To retain the best chromosome set in each generation, an elitism scheme was applied in which only 1 chromosome will be replaced in each generation. For genetic operations, binary tournament selection, single-point crossover and simple mutation are used. The rates of crossover and mutation are 0.5 and 0.1, respectively.

Termination criteria Two termination criteria are defined: criteria (A) is used to stop the entire selection process and criteria (B) is used to stop the fitness evaluation. Both criteria (A) and (B) were set to 5000 and 20000 repetitions, respectively.

2.2 Data Acquisition

For performance evaluation, the acute leukaemia dataset [7] which contains acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML) tumour data are considered. There are 72 samples and 7129 gene expression levels in the dataset. Among 72 sample data, 38 samples (27 ALL, 11 AML) are used for gene selection and the remaining 34 samples (20 ALL, 14 AML) are used for validating the significance of the selected gene subsets. To assess the efficacy of the proposed method, the data normalisation process is ignored.

2.3 Classification

In order to validate the significance of the selected genes, 3 supervised classifiers: MLP, SMO and IBk; are employed using WEKA suite of data mining (http://www.cs.waikato.ac.nz/ml/weka/). The MLP is a 3-layered backpropagation perceptron-based model that employs a sigmoid activation function; the SMO is an implementation of support vector machines (SVMs) that compute the upper bound of support vector weights using a sequential minimal optimisation algorithm; the IBk is an implementation of a k-nearest neighbour (KNN) classifier that employs the similarity distance metric in forming neighbours. All classifiers are implemented with the default settings except for the IBk algorithm where k is set to 3.

3 Results and Discussions

3.1 Gene Selection Results

By repeating the selection process 5000 times using the training set, the top 48 genes based on the selection frequency of at least 50 selections and above are ranked. Table 1 shows the comparison of the selected genes from our method and previous works on the experiment dataset. Out of the 48 selected genes, 28 are consistent with Golub et al. [7]. The significant genes such as CST3 (M27891), zyxin (X95735), c-myb (U22376), adipsin (M84526), CCND3 (M92287), macmarcks (HG1612-HT1612), proteasome IOTA chain (X59417), IL 8 (M28130), azurocidin (M96326) and IL 8 precursor (Y00787) are highly ranked by our method. Amongst these genes, CST3 and zyxin genes have been reported as the strongest predictors in related literature [4, 5, 7, 10, 12].

3.2 Classification Results

As previously described, 34 out of 72 samples are used to assess the significance of the selected genes based on the 3 supervised classifiers. Table 2 shows the classification results based on varying number of selected genes. 33/34 test samples have been correctly classified with an accuracy of 97.06% when there are at least 8 selected genes. The classification performance is reduced when there are less than 8 genes used for class discrimination. Although genes M27891 and X95735 have been identified as the 2 strongest predictors in literature, however, the classification performance based on these combined genes are 91.18%, 85.29% and 91.18% for MLP, SMO and IBk, respectively. By increasing the number of selected genes for classification, the performance has been improved. This confirms that there are many strong predictors that can be used for classification and this supports the similar observations on works [7, 10].

3.3 Related Works

Table 3 shows some related research on the microarray experiment dataset. Several selection techniques have been applied to find the optimal set of genes, including signal-to noise (S2N), Pearson correlations, correspondence analysis (COA), principal component analysis (PCA), between-group/within-group ratio (BSS/WSS) and recursive feature elimination (RFE). Some produce small amount of genes in the selection process and some require more genes for better search performance. However, all existing selection methods based on oligonucleotide microarray data require data preprocessing for optimal search. Depending on the selection approach and classification method, varying data preprocessing techniques are implemented. This has contributes to the genes variability on selection results. Our method, on the other hand, has the ability to extract informative genes without data preprocessing which reduces variation errors on the selected genes.

4 Conclusions

In this paper, a feature selection method based on genetic algorithms and neural networks for oligonucleotide microarray data was developed. For the selection process, the fitness value of the selected gene subset is based on the correctly

Table 1. Comparison of top-48 selected genes by proposed method and previous works. The gene selection is based on the selection frequency of 50 times or more. Genes in *Italic* had been reported by Golub et al. [7].

Rank	Index	Acc No	Description	Marker Genes
1	*1882*	*M27891*	*CST3 Cystatin C*	abcd
2	*4847*	*X95735*	*Zyxin*	acd
3	*5772*	*U22376*	*C-myb gene*	abc
4	*2288*	*M84526*	*Adipsin*	ad
5	*2354*	*M92287*	*CCND3 Cyclin D3*	ab
6	804	HG1612-HT1612	Macmarcks	
7	*4328*	*X59417*	*Proteasome IOTA chain*	ab
8	*6200*	*M28130*	*Interleukin 8 (IL8) gene*	abd
9	*2402*	*M96326*	*Azurocidin gene*	abd
10	*6201*	*Y00787*	*Interleukin 8 precursor*	abd
11	*2121*	*M63138*	*CTSD Cathepsin D*	a
12	1120	J04615	SNRPN Small nuclear ribonucleoprotein polypeptide N	
13	5552	L06797	Probable G protein-coupled receptor LCR1 homolog	b
14	*5501*	*Z15115*	*TOP2B Topoisomerase (DNA) II beta*	b
15	*1704*	*M13792*	*ADA Adenosine deaminase*	b
16	6041	L09209	APLP2 Amyloid beta (A4) precursor-like protein 2	
17	4211	X51521	VIL2 Villin 2	
18	*1928*	*M31303*	*Oncoprotein 18 (Op18) gene*	a
19	4373	X62320	GRN Granulin	
20	*1745*	*M16038*	*LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog*	ac
21	*2642*	*U05259*	*MB-1 gene*	bd
22	6218	M27783	ELA2 Elastatse 2, neutrophil	cd
23	760	D88422	Cystatin A	cd
24	4377	X62654	ME491 gene for Me491/CD63 antigen	
25	*6539*	*X85116*	*Epb72 gene exon 1*	a
26	*3320*	*U50136*	*Leukotriene C4 synthase (LTC4S) gene*	ac
27	1685	M11722	Terminal transferase mRNA	bd
28	*4535*	*X74262*	*Retinoblastoma binding protein P48*	
29	*5039*	*Y12670*	*LEPR Leptin receptor*	ac
30	*5191*	*Z69881*	*Adenosine triphosphatase, calcium*	
31	1779	M19507	MPO Myeloperoxidase	bd
32	*4052*	*X04085*	*Catalase (EC 1.11.1.6) 5'flank and exon 1 mapping to chromosome 11, band p13 (and joined CDS)*	
33	1829	M22960	PPGB Protective protein for beta-galactosidase (galactosialidosis)	
34	6378	M83667	NF-IL6-beta protein mRNA	d
35	*3258*	*U46751*	*Phosphotyrosine independent ligand p62 for the Lck SH2 domain mRNA*	a
36	1133	J04990	Cathepsin G precursor	b
37	*2020*	*M55150*	*FAH Fumarylacetoacetate*	ac
38	*6376*	*M83652*	*PFC Properdin P factor, complement*	d
39	5954	Y00339	CA2 Carbonic anhydrase II	
40	6277	M30703	Amphiregulin (AR) gene	
41	6308	M57731	GRO2 oncogene	d
42	312	D26308	NADPH-flavin reductase	
43	*1249*	*L08246*	*Induced myeloid leukemia cell differentiation protein MCL1*	a
44	*6855*	*M31523*	*TCF3 Transcription factor 3*	
45	*3056*	*U32944*	*Cytoplasmic dynein light chain 1 (hdlc1) mRNA*	
46	878	HG2855-HT2995	Heat Shock Protein, 70 Kda (Gb:Y00371)	
47	*1630*	*L47738*	*Inducible protein mRNA*	
48	1962	M33680	26-kDa cell surface protein TAPA-1 mRNA	b

a represents marker genes reported by Cho and Won [4]
b represents marker genes reported by Culhane et al. [5]
c represents marker genes reported by Li and Yang [10]
d represents marker genes reported by Mao et al. [12]

Table 2. Classification performance comparison based on the number of genes used in the test set.

Number of genes	MLP	SMO	IBk (k=3)
top 2	91.18	85.29	91.18
top 4	94.12	97.06	94.12
top 8	97.06	97.06	97.06
top 16	97.06	97.06	97.06
top 32	97.06	97.06	97.06
top 48	97.06	97.06	97.06

Table 3. Some relevant works on acute leukaemia.

Authors	Selection method	Classification method	Data preprocessing step	Marker genes identified
Proposed method	GANN	-	-	48
Golub et al. [7]	S2N	WV	mean and deviation normalisation	50
Cho and Won [4]	Pearson	ensemble MLPs	max-min normalisation	50
Culhane et al. [5]	COA, PCA	BGA	for COA: negative values transformation; for PCA: mean and deviation normalisation	50
Li and Yang [10]	stepwise selection	logistic regression	log transformation	12
Dudoit et al. [6]	BSS/WSS	various discriminant methods	thresholding, filtering, log transformation, mean and variance normalisation	-
Mao et al. [12]	RFE, F-test	SVMs	preprocessing step used in [6]	20
Lee and Lee [9]	BSS/WSS	SVMs	preprocessing step used in [6]	20

labelled sample data computed by neural networks. For performance evaluation, the selected genes have been evaluated by implementing commonly used classifiers based on a separate set of unknown sample data. The experimental results show that our method is able to identify a set of informative genes without data preprocessing this has reduced the potential of gene variability problems.

References

1. Beiko, R.G., Charlebois, R.L.: GANN: Genetic algorithm neural networks for the detection of conserved combinations of features in DNA. BMC Bioinformatics 6, (2005)
2. Bevilacqua, V., Mastronardi, G., Menolascina, F.: Genetic Algorithm and Neural Network Based Classification in Microarrat Data Analysis with Biological Validity Assessment. ICIC 3, 475–484 (2006)

3. Cho, H.S., Kim, T.S., Wee, J.W., Jeon, S.M., Lee, C.H.: cDNA microarray data based classification of cancers using neural networks and genetic algorithms. Nanotech'03: Nanotechnology Conference and Trade Show, proceedings, (2003)
4. Cho, S.B., Won, H-H.: Cancer classification using ensemble of neural networks with multiple significant gene subsets. Applied Intelligence 26, 243–250 (2007)
5. Culhane, A.C., Perrière, G., Considine, E.C., Cotter, T.G., Higgins, D.G.: Between-group analysis for microarray data. Bioinformatics 18, 1600–1608 (2002)
6. Dudoit, S., Fridlyand, J., Speed, T.P.: Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. Journal of the American Statistical Association 97, 77–87 (2002)
7. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 286, 531–536 (1999)
8. Karzynski, M., Mateos, Á., Herrero J., Dopazo, J.: Using a Genetic Algorithm and a Perceptron forFeature Selection and Supervised Class Learning in DNA Microarray Data. Artif. Intell. Rev. 20, 39–51 (2003)
9. Lee, Y., Lee, C-K.: Classification of multiple cancer types by multicategory support vector machines using gene expression data. Bioinformatics 19, 1132–1139 (2003)
10. Li, W, Yang, Y.: How Many Genes Are Needed for a Discriminant Microarray Data Analysis? CAMDA'00: Critical Assessment of Techniques for Microarray Data Analysis, proceedings, 137–150 (2002)
11. Lin, T-C., Liu, R-S., Chao Y-T., Chen, S-Y.: Multiclass Microarray Data Classification Using GA/ANN Method. PRICAI'06: Trends in Artificial Intelligence, 9th Pacific Rim International Conference on Artificial Intelligence, proceedings 4099, 1037–1041 (2006)
12. Mao, Y., Zhou, X., Pi, D., Sun, Y., Wong, S.T.C.: Multiclass Cancer Classification by Using Fuzzy Support Vector Machine and Binary Decision Tree With Gene Selection. Journal of Biomedicine and Biotechnology 2005, 160–171 (2005)