Evaluation of Computational Classification Methods for Discriminating Human Heart Failure Etiology Based on Gene Expression Data

HY Wang, H Zheng, F Azuaje

School of Computing and Mathematics, University of Ulster, Northern Ireland, UK

Abstract

Human heart failure is a complex syndrome that can be initiated by a variety of clinical conditions and geneticl factors. Gene expression profiling offers opportunities to study changes in transcriptional activity in heart failure samples of different etiologies. This paper evaluates machine and statistical learning models for supporting the identification of heart failure etiology based on gene expression data. Six supervised classification models were evaluated on a publicly-available human heart failure dataset. The Naive Bayes, Support Vector Machines, and k-Nearest Neighbours achieved the most significant prediction performances. Using a correlation coefficient-based gene-ranking criterion, the impact of the number of genes on the prediction performance was investigated. Information from the top 5 genes was sufficient to accurately distinguish between ischemic and idiopathic samples.

1. Introduction

Human heart failure is one of the major causes of morbidity and mortality in most developed countries [1]. It can be initiated by a variety of clinical conditions and genetic factors, such as hypertension, myocardial infraction (MI), and mutations in sarcomeric proteins [2], [3]. Heart failure arising from different etiologies may involve distinct pathophysiological mechanisms, responses to certain pharmacological treatments and prognosis [4], [5]. It is expected that patients with heart failure caused by coronary disease have worse long-term outcomes than other etiology subgroups [4]. Felker *et al.* found that there was a significant interaction between heart failure etiology and the effect of milrinone [5]. It has been suggested that an individualized, etiology-based therapeutic approach could result in major progress in treating heart failure patients [6], [7]. But due to the complex mechanisms involved in heart failure, early and accurate diagnosis remains a difficult challenge [8].

Recent advances in large-scale gene expression profiling techniques offer new opportunities to study human heart failure arising from different etiologies. Such analyses may reveal distinct etiology-specific genomic patterns and support the identification of relevant biomarkers that differentiate different heart failure etiologies. For example, based on the examination of the gene expression of 21 non-ischemic (NICM) cardiomyopathy samples, 10 ischemic (ICM) samples, and 6 normal heart samples, Kittleson *et al.* [9] identified 257 genes differentially expressed in NICM and 72 genes significantly associated with ICM samples. Using statistical techniques to analyze gene expression profiles of samples from 7 normal and 8 failing hearts, Tan *et al.* [10] studied gene expression fingerprints of human heart failure. They found 103 genes (out of 6606) to be differentially expressed between failing and normal heart samples. Recently, Huang *et al.* [7] presented a comparative study of five classification techniques for distinguishing heart failure etiologies using two gene expression datasets generated at two independent laboratories. Using leave-one-out cross-validation, they found that the five statistical methods, including partial least squares, nearest shrunken centroids and random forests, showed similar prediction performances on each of the two datasets.

In this study we evaluated six computational classification models for supporting the identification of human heart failure etiology based on gene expression data. The following questions are addressed: Can machine and statistical learning-based classifiers accurately discriminate between heart failure etiologies solely based on gene expression data? Can we achieve comparable or relatively high prediction performances with a smaller subset of genes? The classification problem was to distinguish between ischemic and idiopathic samples.

The remainder of this paper is organized as follows. Section 2 briefly describes the dataset under study. A description of the prediction models and statistical evaluation techniques is given in Section 3. The results are presented in Section 4. The discussion of results and conclusions, together with future research, are given in

Section 5.

2. Data

A total of 59 samples, including 32 ischemic and 27 idiopathic cardiomyopathy samples, were analyzed. The expression profiles of these samples were measured by Affymetrix Human Genome chips (HG-U133 plus 2) containing 54,675 gene probes. The dataset is freely available at the Program for Genomic Applications website [http://www.cardiogenomics.org], which is a project of the U.S National Heart, Lung and Blood Institute (NHLBI).

3. Methods

3.1. Machine and statistical learning models

Six supervised classification models: Naive Bayes (NB), Support Vector Machines (SVM), Multilayer Perceptron (MLP), k-Nearest Neighbors (KNN), C4.5 Decision Trees and Random Forests (RF) were evaluated. All these models were implemented using the *Weka* package [12]. The implementation of SVM was based on the sequential minimal optimization algorithm developed by Platt [13]. Several models with different learning parameters were implemented. The models reported here used the following learning parameters. The number of learning epochs for MLP was set to 500. The MLP models included one hidden layer with 10 neurons, and an output layer with 2 neurons referring to the two classes: ischemic and idiopathic class. For the C4.5 algorithm, the minimum number of instances per leaf was equal to 2. The number of neighbors in the KNN was set to 5. The number of decision trees to be generated in the RF was equal to 10. A more detailed description of these learning algorithms and their parameters can be found in [12].

3.2. Evaluation and validation

Given that the number of samples under study is limited, a leave-one-out cross-validation procedure was carried out to assess the quality of the prediction models. Each classifier was evaluated based on three statistical measures: *precision* (*Pr*), *specificity* (*Sp*), and *sensitivity* (*Se*). Permutation tests were also implemented to estimate the statistical significance of the prediction performances, i.e. we randomly reshuffled the class labels of the samples several times and generated 1000 permutated datasets. For each permutated dataset, classifiers were implemented again. The statistical significance was then established based on the number of times the permutated datasets produced better results than the original dataset.

3.3. Gene ranking criterion

Based on a well-known correlation coefficient-based ranking criterion [11], each gene, g_i, was ranked in terms of its capacity to distinguish between classes. Let $\mu_1(g_i)$ and $\mu_2(g_i)$ be the mean values of g_i for the classes 1 and 2, $\sigma_1(g_i)$ and $\sigma_2(g_i)$ be the standard deviations of g_i for the classes 1 and 2, the gene ranking score, S_i, can be calculated as:

$$S_i = \frac{\mu_1(g_i) - \mu_2(g_i)}{\sigma_1(g_i) + \sigma_2(g_i)} \quad (1)$$

Large values of $|S_i|$ indicate a stronger correlation between the expression of a gene, g_i, and one of the classes under consideration.

4. Results

In order to study the effect of the number of genes on the classification performance, all the genes were ranked and a selected subset of top ranked genes were used to train the classifiers. Figure 1 shows the classification accuracy of the classifiers with different numbers of top ranked genes ranging from 1 to 100.

A closer examination of the results presented in Figure 1 reveals that:

1. The impact of the number of genes on the prediction results is dependent on the classification model adopted. Some classifiers are more sensitive to the number of input genes than others. For example, the C4.5-based model exhibits a relatively large variation of prediction accuracy that ranges from 78.0% to 91.5%. Moreover, there is no clear relationship between the number of genes and the performance of C4.5. In the case of NB and SVM, the number of genes used as inputs to the model may not have a significant impact on the prediction results when the number of input genes is greater than 5.
2. High classification accuracy can be obtained by using only a small subset of genes. For example, both SVM and KNN achieve the best prediction accuracy (94.9%) when they used the top 5 genes (*RPS10*, *PLN*, *BCL2L1*, *226203_at*, and *217024_x_at*) as model inputs. This indicates that a relatively small, selected subset of genes may be sufficient to accurately distinguish between the classes studied here.
3. Using the top 5 ranked genes, all the 6 classifiers can achieve relatively high prediction performances in

terms of accuracy, sensitivity and specificity, as shown in Table 1. Prediction accuracies between 89% and 94% were observed. This suggests that these 5 genes may play a significant role in distinguishing between ischemic and idiopathic cardiomyopathy patients. Surprisingly, only the top 3 genes, as described in Table 2, have functional annotations in the Gene Ontology database [14].

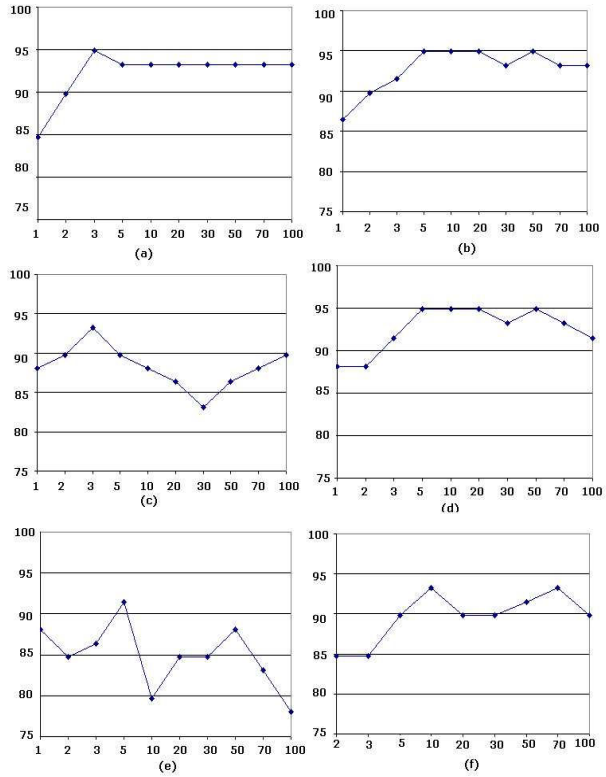

Figure 1 The classification accuracy (shown on Y axis) against the number of genes (shown on X-axis) for each classifier: (a) NB; (b) SVM; (c) MLP; (d) KNN; (e) C4.5; (f) RF

To estimate the statistical significance of the proposed computational classification methods, a 1000-runs permutation test was carried out. In 999 out of 1000 permutated datasets, the results were significantly worse than the results obtained using the original data in terms of *Ac*, *Se*, and *Sp*. For example, when implementing the permutation test for the NB classifier with the top 5 genes, the obtained average values of *Ac* (48.9%, for Ischemic Class) were significantly lower than the results shown in Table 1 (p = 0.001), which indicates that the high classification accuracy values shown in Table 1 are unlikely to have been obtained by chance. Similar predictive responses were obtained from the permutation tests for the other classifiers.

Table 1 Prediction results for different classifiers using leave-one-out cross validation with the top 5 genes: *RPS10*, *PLN*, *BCL2L1*, *226203_at*, and *217024_x_at*.

Model	*Ac* (%)	Ischemic Class			Idiopathic Class		
		Pr (%)	*Se* (%)	*Sp* (%)	*Pr* (%)	*Se* (%)	*Sp* (%)
NB	93.2	91.2	96.9	88.9	96.0	88.9	96.9
SVM	94.9	91.4	100	88.9	100	88.9	100
MLP	89.8	90.6	90.6	88.9	88.9	88.9	90.6
KNN	94.9	91.4	100	88.9	100	88.9	100
C4.5	91.5	90.9	93.8	88.9	92.3	88.9	93.8
RF	89.8	88.2	93.8	85.2	92.0	85.2	93.8

Table 2 Description of top 3 genes. Biological process annotations for each gene were obtained with GenNav tool (http://mor.nlm.nih.gov/perl/gennav.pl)

Rank	Ranking Score	Gene Symbol	Biological process annotations
1	1.09	RPS10	• Protein biosynthesis
2	1.05	PLN	• Muscle contraction • Circulation
3	1.01	BCL2L1	• Anti-apoptosis • Negative regulation of survival gene product activity • Apoptotic mitochondrial changes

5. Discussion and conclusions

Previous research has shown that the cause of heart failure may affect the response to drug treatment and long-term prognosis [6]. Based on a publicly-available heart failure gene expression dataset, this paper evaluated machine and statistical predictive models for supporting the identification of heart failure etiology. The predictive performances of 6 different supervised classification

models were compared. NB, SVM and KNN achieved the most significant prediction performances (above 94% of classification accuracy). Using a correlation coefficient-based gene ranking criterion, the influence of the number of genes on the predictive performances was investigated. The results indicate that by incorporating only the top 5 genes, all the 6 classification models may achieve relatively high prediction performances with classification accuracies ranging from 90% to 95.9%. The continual increase of the number of genes does not significantly contribute to the improvement of classification performance, especially in the case of NB and SVM-based models.

The gene-ranking criterion used in this paper is by no means the most optimal technique to estimate gene relevance. The application of other feature ranking techniques, such as *F*-statistics [7], deserves further investigation. However, this study demonstrated that it is feasible to detect a small set of top-ranked genes that can accurately distinguish between heart failure classes. Also these genes may motivate additional computational and experimental studies to assess their relevance in heart failure pathways and as potential drug targets. The application of more advanced gene ranking techniques, such as the maximum-relevance-minimum-redundancy-based gene selection method, [15] is an important task of future research.

The techniques assessed in this paper, as well as future investigations, will contribute to a European Union Sixth Framework Programme (FP6) project, which aims to detect potential drug targets relevant to heart failure and atherosclerosis.

Acknowledgements

This work was supported in part by a grant from EU FP6, CARDIOWORKBENCH project, to FA.

References

[1] Hobbs FDR. Management of heart failure: evidence versus practice. Does current prescribing provide optimal treatment for heart failure patients? British Journal of general Practice 2000, 50: 735-742.

[2] Levy D, Larson MG, Vasan RS, Kannel WB, Ho KK. The progression from hypertension to congestive heart failure. JAMA 1996, 275:1557-1562.

[3] Nicol RL, Frey N, Olson EN. From the sarcomere to the nucleus: Role of Genetics and Signaling in Structural Heart Disease. Annual Review of Genomics and Human Genetics 2000; 1: 179-223.

[4] Cleland JGF, Swedberg k, Poole-Wilson PA. Successes and failures of current treatment of heart failure. The Lancet 1998, 352(suppl): 19-28.

[5] Felker GM, Benza RL, Chandler AB, Leimberger JD, Cuffe MS, Califf RM, Gheorghiade M, O'Connor CM. Heart failure etiology and response to milrinone in decompensated heart failure. Journal of the American College of Cardiology 2003, 41: 997-1003.

[6] Follath F, Cleland JG, Klein W, Murphy R. Etiology and response to drug treatment in heart failure. Journal of the American College of Cardiology 1998, 32: 1167-1172.

[7] Huang X, Pan W, Grindle S, Han X, Chen Y, Park SJ, Miller LW, Hall J. A comparative study of discriminating human heart failure etiology using gene expression profiles. BMC Bioinformatics 2005 6: 205.

[8] NHS, The diagnosis and drug treatment of heart failure. MeReC briefing 2001, 15:1-8.

[9] Kittleson MM, Minhas KM, Irizarry RA, Ye SQ, Edness G, Breton E, Conte JV, Tomaselli G, Garcia JGN, Hare JM. Gene expression analysis of ischemic and nonischemic cardiomyopathy: Shared and distinct genes in the development of heart failure. Physiological Genomics 2005, 21: 299-307.

[10] Tan F, Moravec CS, Li J, Apperson-Hansen C, McCarthy PM, Young JB, Bond M. The gene expression fingerprint of human heart failure. PNAS 2002, 99: 11387-11392.

[11] Golub TR, Slonim DK, Tamayo P, Huard C, Gassenbeck M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999, 286: 531-537.

[12] Witten IH, Frank E: Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[13] Platt J. Fast Training of Support Vector Machines using Sequential Minimal Optimization. Advances in Kernel Methods - Support Vector Learning, B. Schoelkopf, C. Burges, and A. Smola, eds., MIT Press, 1998.

[14] The Gene Ontology Consortium. Creating the gene ontology resource: Design and implementation. Genome Research 2001, 11: 1425-1433.

[15] Ding C and Peng H. Minimum redundancy feature selection from microarray gene expression data. Journal of Bioinformatics and Computational Biology 2005, 3: 185-205.

Address for correspondence.

Name: Haiying Wang
Full postal address: School of Computing and Mathematics, University of Ulster at Jordanstown, BT37 0QB, UK
E-mail address: hy.wang@ulster.ac.uk