Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2012) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

Complex Adaptive Systems, Publication 2
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2012- Washington D.C.

GRNN Ensemble Classifier for Lung Cancer Prognosis Using Only Demographic and TNM features

J. David Schaffer[a], Jin Woo Park[a], Erin Barnes[a], Qiyi Lu[b], Xingye Qiao[b], Youping Deng[c], Yan Li[c], Walker H. Land, Jr.*[a]

[a] *Dept. of Bioengineering, Binghamton University, 85 Murry hill Road, Vestal ,NY, 13902-6000, USA*
[b] *Dept. of Mathematical Sciences, Binghamton University, PO Box 6000, Binghamton ,NY, 13902-6000, USA*
[c] *Rush University Cancer Center, Department of Internal Medicine, Kidston House, 630 S. Hermitage Avenue, Chicago, IL, 60612, USA*

Abstract

Predicting the recurrence of non-small cell lung cancer remains a clinical challenge. The current best practice employs heuristic decisions based on the TNM classification scheme that many believe can be improved upon. Much research has recently been devoted to searching for gene signatures derived from gene expression microarrays for this challenge, but a consensus signature is still elusive. We present an approach to first create a benchmark for recurrence prediction based only upon gender, age and TNM features that uses several learning classifier induction methods and combines them into an ensemble using a recent extension of the general regression neural network. Using this approach on a pooled sample of 422 patients from two previously published studies (Shedden and Raponi), we demonstrate error rates in the low 20% for both false positives and negatives. Future work will focus on discovering if gene signatures can be discovered that can improve this performance.

Keywords:Bioinformatics; Biomedical; non-small cell lung cancer; recurrence prediction, linear regression, support vector machine; generalized regression neural networks; probabilistic neural networks

1. Background

Lung cancer is the leading cause of death in cancer patients worldwide. In 2011, the American Cancer Society predicted that 156,940 people would fall victim to the disease, accounting for 27% of all cancer deaths [1]. Of the 221,130 estimated cases that would be diagnosed in 2011, 85% would have late stage tumors (II, III, IV) that have begun to advance. Due primarily to this late stage diagnosis, the 5-year survival rate of lung cancer patients is 16%. For these patients, treatment often includes surgical resection of tumors if possible, post-operative radiation, and adjuvant chemotherapy. The 5-year survival rate for early stage (I) non-small cell lung cancer (NSCLC) patients is 53% [1] and treatment at this stage often only includes surgical resection [2]. However, 35%-50% of these patients will suffer a relapse of the disease within 5 years of surgery [5]. As a result, many doctors resort to administering post-operative chemotherapy, which in most cases improves survival, but this approach is controversial, expensive, and may not be necessary for all patients. Doctors currently lack a validated and clinically accepted method to predict which patients are at a high risk of recurring cancer [4]. Those patients that are at a high risk of recurrence might benefit from post-operative adjuvant chemotherapy, whereas those patients that are at a low risk can be spared the side effects of chemotherapy [5]. Much research is currently focused on discovering gene expression signatures that seek to improve our ability to identify patients at high risk of recurrence [6]. The work reported in this paper is a first step in this direction by determining how well methods based on statistical learning theory can do when given only simple demographic and TNM (Tumor, Node, Metastasis) features in a data set that combines information from two publically available investigations of non-small cell lung cancer [3, 5]. With this as a benchmark, future work will investigate to what extent gene expression information can further improve recurrence prediction.

2. Methods

2.1 Data Set

The data set consisted of 422 patients drawn from two publically available NSCLC studies: Raponi et al. [3] (n=130) and Shedden et al. [5] (n=292). Work is on-going to compile and pool the microarray data from these studies with the aim of searching for robust gene signatures, but herein we explore only demographic and TNM (Tumor, lymph Node, Metastases) features. The features comprised gender, age, tumor stage (IA, IB, IIA, IIB, IIIA, IIIB), cancer type (adenocarcinoma, squamous cell), tumor size (grade 0-4), lymph node involvement (grade 0-3), metastasis (y/n), and stage grouping (codes 0-3). The outcome measure to be predicted was good prognosis (recurrence or death > 36 months) or poor prognosis (≤ 36 months). All categorical variables were mapped linearly to the range [0,1], while age was mapped using the *tanh* (hyperbolic tangent) function setting the minimum age in the data to 0.1 and the maximum to 0.9.

We broke the data set into five folds stratified so that each fold contained roughly equal proportions of both cancer cell types and prognosis categories. Each classifier induction method (see below) was applied to each set of four folds and tested on the fifth. The resulting predictions on the left-out fold were concatenated and compared to the ground truth prognosis to compute an AUC (area under the ROC: receiver operator characteristic curve) and errors. Finally, we use the Generalized Regression Neural Networks (GRNN) “oracle” to combine the three classifiers into an ensemble.

2.2 Linear Regression

We used the R system (version 2.13.0), linear regression of the independent variables against “prognosis”.

2.3 Support Vector Machine (SVM)

The support vector machine (SVM) that was used was the libsvm learner in Orange [7,8]. A simple grid search was done to find the best performing parameters and the kernel that was used in the SVM was the radial basis function with a gamma value of 0.00010. This value gave well performing SVMs for all five folds.

2.4 Probabilistic Neural Network (PNN)

The PNN is a classifier that makes very effective use of the shapes of population distributions. Given a sample/ case/control that has previously not been classified, the PNN determines the probability that the point belongs to a particular class based on populations of points whose class identity is known *a priori*. The PNN makes this determination in two steps.

It uses the multivariate Parzen density estimator to estimate (a constant multiple of) the density function of each population at the known point. Then it computes the Bayesian posterior probability of membership in each population and assigns the unknown to the class having the highest posterior probability. In particular, suppose we have a training set composed of *n* cases. Each case *i* (*i*=1, …, *n*) consists of:

$$x_{i,j} \quad j=1, \ldots, p$$

These *predictor variables* determine the relative efficacy of the prediction models. The observed values of the predictor variables for the unknown case are:

$$x_j \quad j=1, \ldots, p$$

These are the values of the observed *gate variables*. The weighted Euclidean distance function is often employed:

$$D(X,X_i)=\sum_{j=1}^{p}\left(\frac{x_j-x_{i,j}}{\sigma_j}\right)^2 \qquad (1)$$

The (unnormalized) density function is then given by Parzen's formula:

$$g(x)=\frac{1}{n}\sum_{i=1}^{n}\exp(-D(x,x_i)) \qquad (2)$$

The above procedure is repeated for each of the *K* classes, giving the unnormalized density $g_k(x)$ for k=1, …, *K*. The Bayesian probability that the unknown case was drawn from class *k* is as follows:

$$\frac{g_k(x)}{\sum_{i=1}^{K} g_i(x)} \qquad (3)$$

From the above formulation, the PNN has three interesting properties. **First**, under reasonable assumptions, and as the training set size increases, the PNN is an asymptotically optimal classifier. **Second**, the PNN is fully nonlinear. It does not impose constraints of linearity on the model (as does ordinary discriminate analysis), nor does it require linear separability like many other classifiers. **Finally**, the PNN has the ability to compute Bayesian probabilities (including the use of priors if desired), which is a great advantage in many practical applications.

2.5 GRNN Ensemble Formulation Summary

The objective is to design an ensemble processor that uses the gate variables to intelligently combine the outputs of the competing models [9]. A background and history of ensemble processing may be found in Land et al. [10]. Once the expected error of each prediction model is estimated, these expected errors are used to compute the weights for each model. When an unknown case is processed, the gate variables are used by the GRNN to decide which models are likely to be best for this particular case. These models are weighted more heavily than the likely inferior models. In particular, one has a training set composed of *n* cases. Each case *i* (*i* =1, …, *n*) consists of *p* gate variables: $x_{i,j}$ where *j* =1, …, *p*. These *gate variables* determine in some way the relative efficacy of the prediction models. The *m* competing prediction models provide outputs $q_{i,k}$ for each case *i* where *k* =1, …, *m*. The desired output (the target value) for case *i* is y_i.

For the gate variables and model outputs of a trial case that is to be evaluated, just one subscript is used : x_j where *j* =1, …, *p*, are the values of the observed *gate variables*, and q_k where *k*=1, …, *m*, are the computed outputs from the *m* competing prediction models for this new case.

Define the weighted Euclidean distance (as determined by the gate variables) between training case, *i*, and the trial case. Then the GRNN oracle’s predicted squared error for model *k* may be shown to be:

$$\hat{e}_k(\mathbf{x})=\frac{\sum_{i=1}^{n}(y_i-q_{i,k})^2\exp(-D(\mathbf{x},\mathbf{x}_i))}{\sum_{i=1}^{n}\exp(-D(\mathbf{x},\mathbf{x}_i))} \qquad (4)$$

It is desired that the final prediction be a linear combination of the outputs of the competing models:

$$\hat{y} = \sum_{k=1} w_k q_k \qquad (5)$$

Here w_k are the weights for the outputs. If the models have the (desirable) property that their predictions are unbiased, then the following condition is imposed:

$$\sum_{k=1}^{\cdots} w_i = 1 \qquad (6)$$

The linear combination of unbiased estimators having minimum mean-squared error uses weights proportional to the reciprocal of each estimator's variance. If the predicted squared error is used in place of the variance, the following formula is derived for the weights:

$$w_k = \frac{1/\hat{e}_k}{\sum_{l=1}^{m} 1/\hat{e}_l} \qquad (7)$$

The GRNN ensemble processor is trained (i.e., the *p* sigma weights in the weighted Euclidean distance in (1) are optimized) in the leave-one-out validation manner. Differential evolution [11] was used to optimize the sigma weights over the set of training cases.

3. Results

Table 1 shows the correlations among the predictions made by each classifier among the 422 cases. The modest correlations with the ground truth (prognosis class) and the generally modest correlations among the classifier predictions, suggest that the GRNN oracle might be able to learn an ensemble classifier that would do better than any of the three alone.

Table 1. Correlations among classifier predictions and the ground truth

	Regression	SVM	PNN	prognosis
Regression	1			
SVM	0.911	1		
PNN	0.700	0.701	1	
Prognosis	0.607	0.564	0.488	1

Table 2 shows the AUC (area under the ROC curve) and error counts for false positives (FP) and false negatives (FN) using a threshold of 0.5 (i.e. any prediction above 0.5 is called a positive, below a negative classification). The ensemble's error rates (Table 2) coincide with the SVM's because the SVM tends to create a substantial gap between the positive and negative classes, and so is least sensitive among the chosen classifiers to the choice of threshold. Perhaps curiously, the SVM also yielded the worst AUC.

Table 2. Results from 5-fold cross validation of individual classifier and the ensemble

Classifier	AUC test	CV errors FP Test set		CV errors FN Test set		CV total errors	
Regression	0.830	50/177	28.2%	39/245	15.9%	89/422	21.1%
SVM	0.783	38/177	21.5%	55/245	22.4%	93/422	22.0%
PNN	0.822	22/177	12.4%	116/245	47.3%	138/422	32.7%
GRNN	0.847	38/177	21.5%	55/245	22.4%	93/422	22.0%

The ROC curves are shown in Figure 1. From these and Table 2, we see that the GRNN ensemble makes a modest improvement on the individual classifiers. All tend to converge near the middle point. Above this threshold (i.e. preferring sensitivity to specificity), the linear regression classifier performs best among the individuals, and below (i.e. preferring specificity to sensitivity), the PNN performs best, while the ensemble provides the best across the board.

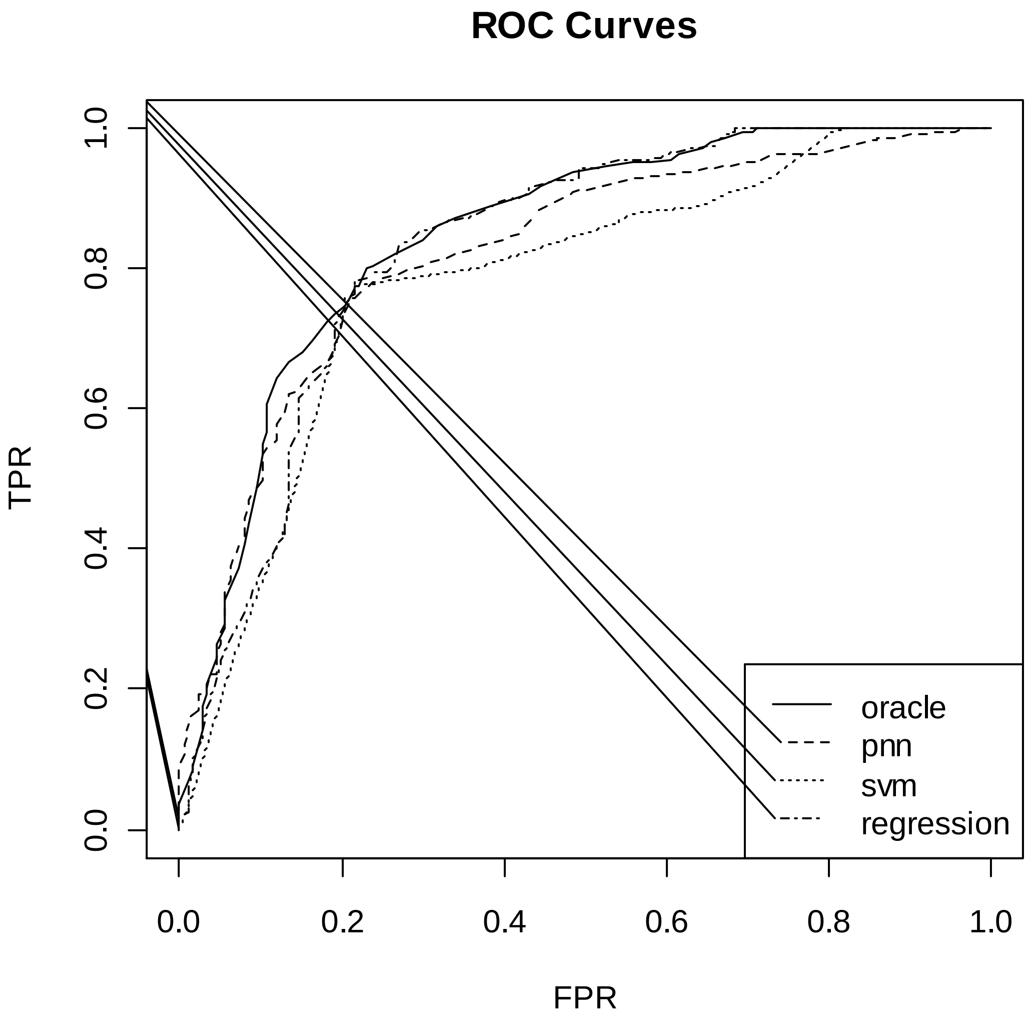

Fig. 1. Oracle and PNN, SVM, regression ROC Curves

4. Conclusions

This work provides a benchmark for predicting lung cancer recurrence within three years using only gender, age and TNM codes that are readily available in today's clinical practice. Using leave-one-out cross validation, the GRNN ensemble classifier seems capable of error rates in the low 20% range for both false positives and false negatives. The data comprised 422 patients for whom gene expression microarray data are also avalable so the possibility of improved precision by adding gene signatures can be exploered. The 36-month threshold for recurrence was chosen because the cohorts were well balanced. We also plan to examine 60 months, as this may be a more clinically meaningful thershold.

Work is still ongoing to prepare the microarray data for analysis: correcting for possible center effects (distributions of microarray values being different for the different research centers), and coarse feature reduction using significance analysis for microarays (SAM). Once the full data set is assembled, we plan to explore for gene signatures based upon microarray measurements that can be combined with these clinical measures to make further increases in the accuracy of recurrence prediction. Of course, any potentially useful findings will need independent verification in larger cohorts.

References

1. Cancer facts & figures. American Cancer Society 2011, 15-16.
2. Lu Y, Lemon W, Liu PYY, Yi Y, Morrison C, Yang P, Sun Z, Szoke J, Gerald WL, Watson M, Govindan R, You M: A gene expression signature predicts survival of patients with stage I non-small cell lung cancer. PloS medicine 2006, 3(12).
3. Raponi M, Zhang Y, Yu J, Chen G, Lee G, Taylor JMG, Macdonald J, Thomas D, Moskaluk C, Wang Y, Beer DG: Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. Cancer research 2006, 66(15):7466-72[http://cancerres.aacrjournals.org/ cgi/content/abstract/66/15/7466].
4. Subramanian J, Simon R: Gene expression-based prognostic signatures in lung cancer: ready for clinical use? Journal of the National Cancer Institute 2010, 102(7):464-474.
5. Shedden K, Taylor JMG, Enkemann SA, Tsao MS, Yeatman TJ, Gerald WL, Eschrich S, Jurisica I, Giordano TJ, Misek DE, Chang AC, Zhu CQ, Strumpf D, Hanash S, Shepherd FA, Ding K, Seymour L, Naoki K, Pennell N, Weir B, Verhaak R, Ladd-Acosta C, Golub T, Gruidl M, Sharma A, Szoke J, Zakowski M, Rusch V, Kris M, Viale A, Motoi N, Travis W, Conley B, Seshan VE, Meyerson M, Kuick R, Dobbin KK, Lively T, Jacobson JW, Beer DG: Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nature medicine 2008, 14(8):822-7.
6. Lin J, Beer, DG: Molecular Predictors of Prognosis in Lung Cancer, Annals of Surgical Oncology, 2012 Feb;19(2):669-76. Epub 2011 Aug 6.
7. Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
8. Tomaž Curk, Janez Demšar, Qikai Xu, Gregor Leban, Uroš Petrovič, Ivan Bratko, Gad Shaulsky, Blaž Zupan. Microarray data mining with visual programming. Bioinformatics. 2005 Feb 1;21(3):396-8.
9.W. Land, L. Wong, D.W. McKee, T.Masters, F.R. Anderson, " Breast Cancer Computed Aided Diagnosis (CAD) Using a Recently Developed SVM/GRNN Oracle Hybrid"Paper number 425, for SMCC'03, 0-7803-7952-7-7/03/S17,2003
10. Land Jr., W.H., Margolis, D., Kallergi, M. and Heine, J.J. (2010) 'A kernel approach forensemble decision combinations with two-view mammography applications',*Int. J. Functional Informatics and Personalised Medicine*, Vol.2, 2010.
11. K. Price, and R. Storn, "Differential Evolution", *Dr. Dobb's Journal*, pp. 18-24, April, 1997.