Improving Clinical Relevance in Ensemble Support Vector Machine Models of Radiation Pneumonitis Risk

Todd W. Schiller and Yixin Chen
Washington University in St. Louis
St. Louis, Missouri, USA
tschiller@acm.org, chen@cse.wustl.edu

Issam El Naqa and Joseph O. Deasy
Washington University School of Medicine
St. Louis, Missouri, USA
elnaqa@wustl.edu, jdeasy@radonc.wustl.edu

Abstract

Patients undergoing thoracic radiation therapy can develop radiation pneumonitis (RP), a potentially fatal inflammation of the lungs. Support vector machines (SVMs), a statistical machine learning method, have recently been used to build binary-outcome RP prediction models with promising results. In this work, we (1) introduce a feature-ranking selection step to improve the parsimony of our previous ensemble SVM model (2) show that ensembles of SVMs provide a statistically significant performance improvement in the area under the cross-validated receiver operating curve and (3) apply Platt's tuning to the component SVMs to generate probability estimates in order to augment clinical relevance.

1. Introduction

Radiation pneumonitis (RP) is a potentially fatal inflammation of the lungs that can result from thoracic radiation therapy. Numerous factors, such as maximum dose [9] and gender [16, 3], have been shown to correspond RP incidence. A tabulated summary of previous findings can be found in Table IV of Das et al.'s work in [3]. There is no clear consensus on a core set of factors affecting RP risk; the lack of consensus can be partly attributed to salient differences across studies including patient populations [5] and model evaluation metrics.

Within the last 5 years, there has been a push to move beyond correlation analysis to the construction of predictive models using machine learning techniques. One such technique relies on SVMs – a class of statistical learning methods. Within an SVM, the input data are mapped into a higher, possibly infinite, dimensional space. The hyperplane best separating the two classes in this feature space is used to define a decision function. The best hyperplane maximizes the margin (distance) between the plane and the closest instances on either side (see Fig. 1).

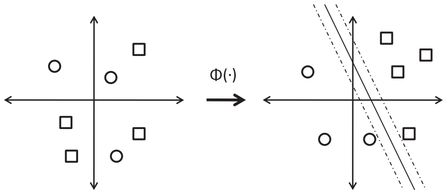

Figure 1. SVM classification: two classes of instances are mapped to an implicit space in which they are separable.

The model's decision function score can be used as a relative indication of risk/certainty – a premise used when calculating the area under the curve (AUC) for a receiver operating characteristic (ROC) curve. The clinical meaning of the difference between scores is not well-defined, however. For instance, a patient with a decision score 20% higher than that of a another patient does not necessarily have 20% greater chance of developing RP. In this way, decision function scores are of limited use in a clinical setting.

Up until now, SVM-only models of RP risk have been binary-outcome – predicting that the patient will either develop or not develop RP. However, support vector machine theory is now sufficiently advanced to correctly produce probability estimates from decision function scores [15, 12].

In [17], we presented a model that fused the output from multiple SVMs to produce an improved binary-outcome model of RP risk. In this paper, we:

1. Introduce a feature-ranking selection step to our previous ensemble method to improve model parsimony

2. Show increased ensemble size provides a statistically significant benefit to model AUC

3. Probabilistically tune component SVM output to im-

prove clinical relevance

These innovations produce a better SVM-based approach to assessing radiation pneumonitis risk and help to characterize challenges in the problem domain.

In the next section, we provide background information on SVM model building, model evaluation, and tuning SVM output to produce probabilistic estimates. In Section 3, we survey related work. In Section 4, we outline our improved ensemble SVM methodology. Results are presented and discussed in Section 5. Finally, we offer concluding remarks in Section 6.

2. Training and evaluating support vector machines

This section briefly introduces SVM training methodology, the cross-validated AUC method for model evaluation, and Platt's method for producing probabilistic outputs from an SVM.

2.1. Support vector machine training

SVMs are trained by finding the hyperplane that best separates the classes in the feature space. The instances are implicitly mapped into the space using a kernel function such as the Gaussian Radial Basis Function (RBF):

$$K_\sigma(x,y) = \exp\left(-\sum_i \frac{(x_i - y_i)^2}{2\sigma_i^2}\right) , \tag{1}$$

where σ is a vector of scaling factors.

Finding the optimal hyperplane can be formulated as an optimization problem:

$$\max_\alpha \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{2}$$

subject to:

$$\sum_i \alpha_i y_i = 0$$
$$\forall i, a_i \geq 0 \; .$$

Finding the optimal α results in a decision function of the form:

$$f(x) = \sum_i \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \; . \tag{3}$$

When the data are not separable in the feature space, a complexity parameter C is introduced to allow training error. C can be included in the model as an extension of the kernel during training:

$$\mathbf{K} \leftarrow \mathbf{K} + \frac{1}{C}\mathbf{I} \; , \tag{4}$$

where $\mathbf{I}$ is the identity matrix [1].

Kernel parameter σ and model parameter C are often selected prior to model building using grid-search [11].The optimization problem in Equation 2 can then be solved using Platt's sequential minimal optimization (SMO) method [14]. Chapelle et al. present an alternative method in which model/parameters are selected concurrently with model building. Alternating SVM training steps and gradient descent parameter selection steps are used to minimize an estimate of generalization error [1].

2.2. Cross-validation analysis

To properly evaluate a model's predictive ability, the training and testing data sets should be disjoint. Data scarcity, however, makes utilizing a separate monolithic validation set undesirable. Instead, cross-validation, a method for alternately using data for training and testing is used. In k-folds cross-validation analysis, the dataset is segmented into k pair-wise disjoint subsets. Each subset is used as a validation set exactly once as the remaining subsets are used to build the model. The results from testing on the k subsets are then combined. When the number of folds is equal to the number data instances (each subset contains one instance), the method is called the leave-one-out (LOO) method.

2.3. Area under the receiver operating characteristic curve

The area under the curve (AUC) for the receiver operating characteristic (ROC) curve is a popular single-value metric of model performance. The ROC is a plot of a model's sensitivity against (1 - specificity) as the decision function threshold is varied, where sensitivity and specificity are defined as:

$$\begin{aligned} \text{sensitivity} &= \frac{\text{\# true positives}}{\text{\# true positives} + \text{\# false negatives}} \\ \text{specificity} &= \frac{\text{\# true negatives}}{\text{\# true negatives} + \text{\# false positives}} \; . \end{aligned}$$

For the radiation pneumonitis problem, the AUC can be interpreted as the probability that a randomly chosen patient that develops RP will be given a higher risk estimate by the model than a randomly chosen patient that does not develop RP [6]. An AUC of 0.5 corresponds to a model that produces random risk estimates, while an AUC of 1.0 corresponds to a perfect model.

Instead of explicitly finding the area under the ROC curve, the AUC can be calculated as:

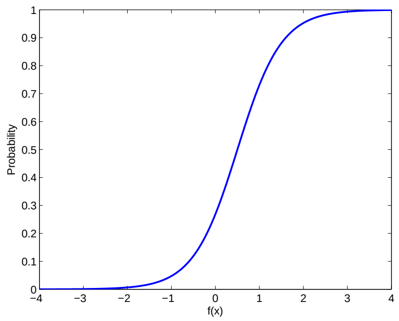

Figure 2. Sigmoid probability curve with A=-2 and B=1

$$\hat{A} = \frac{S_0 - n_0(n_0+1)/2}{n_0 n_1} \, , \tag{5}$$

where S_0 is the rank sum of the positive instances when the decision scores are sorted in ascending order, n_0 in the number of positive instances, and n_1 is the number of negative instances [8].

2.4. Platt's method for probabilistic support vector machine output

The unthresholded SVM decision function produces a real-valued output corresponding to the distance between the instance and the separating hyperplane in the SVM's implicit space. While relative distance to the hyperplane is used as a proxy for relative risk when calculating AUC, the SVM decision function score cannot be used directly as an absolute probability estimate.

Platt offers a relatively simple, but effective, way to convert the decision function score to a probability measure by fitting a sigmoid function of the form

$$P(y=1|f) = \frac{1}{1+\exp(Af+B)} \tag{6}$$

to the SVM output [15]. See Fig. 2 for an example curve with $A = -2$ and $B = 1$.

Let N_+ and N_- be the number of RP positive and negative instances in a training set, respectively. Then the target probabilities for t_+ for positive instances and t_- for RP negative instances are defined as:

$$\begin{aligned} t_+ &= \frac{N_+ + 1}{N_+ + 2} \\ t_- &= \frac{1}{N_- + 2} \, . \end{aligned} \tag{7}$$

The sigmoid parameters A and B are selected by minimizing the cross-entropy error on training data:

$$\min_{A,B} \sum_i t_i \log(p_i) + (1-t_i)\log(1-p_i) \, , \tag{8}$$

where $p_i = P(y_i = 1|f_i)$ and $t_i = t_+$ when instance i is RP positive.

Lin et al. provide pseudo-code for a corrected (and improved) implementation of Platt's method in [12].

3. Related Work

Chen et al. investigate two classes of binary SVM models for significant RP events (2+ grade) in lung cancer receiving 3-D conformal radiotherapy [2]. The first class only includes dosimetric parameters, such as equivalent uniform dose (EUD), while the second also includes clinical parameters – race, age, etc. The classes are evaluated using a 10-fold AUC. Parameter and feature selection is performed within each of the 10-folds. A published model reports both the SVM decision function score and the number of patients in the original dataset that received a higher score given a novel patient/treatment plan. The authors do not formally discuss/investigate the latter rank as an estimation of radiation pneumonitis risk.

El Naqa et al. briefly compare recursive feature elimination (RFE) and logistic regression for feature selection when modeling RP outcomes with an SVM. An SVM with a RBF kernel is constructed using features selected from dosimetric and non-dose variables. The resulting models are compared using Matthew's correlation coefficient (MCC), a function of the confusion matrix for some test set [13].

Other research performed by the same research groups explore real-valued models (analog) models of RP risk. Das et al. extend their SVM investigation in [2] by including the binary SVM model in a model that includes a feed-forward neural network, a decision tree, and a self-organizing map [3]. The models are combined (fused) by taking the mean of 100 binary cross-folded predictions from each of the four models. An extreme output of 1.0 – produced by 400 model positive RP predictions – implies consensus that the patient will suffer RP. The mean is described as a proxy for the probability of a RP event. However, its validity as such is not formally established. Equivalent uniform dose, pre-radiotherapy chemotherapy, and gender are chosen as variables for a logistic regression of the fusion function probabilities. The fit of the regression is demonstrated graphically.

Hope et al. construct a 3-variable logistic model of radiation pneumonitis using features selected via statistical bootstrapping. Though their method does not use SVMs, their method of model comparison is notable. Patients are binned into 6 risk groups according to predicted RP risk values. The average predicted risk value within each risk bin is compared graphically to the actual incidence of RP experienced by patients within the bin [9].

4. Methods

This section briefly outlines our ensemble method in [17] and provides implementation details for the methods specific to this work. All the methods were implemented in Matlab 7.8.0 (R2009a).

4.1. Data set description

The data set is composed of 209 patients that underwent radiation treatment for non-small-cell lung cancer between 1991 and 2001. Data for each patient include clinical, treatment, and tumor location factors such as age, gender, performance status (overall patient health), the maximum dose to the heart, the lateral position of the tumor (COMLAT), and the superior-inferior position of the tumor (COMSI). Each feature is scaled to the range [0,1], inclusive. Patients that developed WUSTL Grade 2+ and RTOG Grade 3+ RP events were labeled as RP positive (a summary of grading systems can be found in Table 1 of [9]). Using this standard, 48 (23%) patients were considered to have exhibited clinically significant RP. A detailed description of the data set can be found in [4].

4.2. Ensemble of support vector machines

Instead of the single SVM approach used by Chen et al. [2], we use an ensemble of SVMs to address data imbalance and exploit potential synergies [17, 3]. As in our previous work in [17], the data is randomly partitioned into equally-balanced subsets. Each of these partitions is used as the underlying training data for an SVM with a Gaussian RBF kernel. The decision function for the ensemble classifier is the mean of the decision function scores of the component classifiers. Each component SVM is built using Chapelle et al.'s method mentioned in Section 2.1. The method is used to minimize a support vector span estimate of the LOO error [18]. It is important to re-emphasize that model parameter C and kernel parameter σ are selected for each SVM during model building, as opposed to separately before.

4.3. SVM feature selection

Features are selected according to a modified version of the AUC-maximizing forward selection algorithm in [2]. As with component SVM construction, training data is randomly partitioned into equally-balanced subsets to be used as underlying data for a larger set of feature selection SVMs. For each of these SVMs, features are added / randomly substituted into the model until the 10-fold cross-validated AUC for the SVM fails to improve. To maintain model parsimony and limit training time, the maximum number of features selected by each classifier is limited to five. The feature selections are compiled to rank the features according to the number of times each feature was selected. The set of top-ranked features are used as the feature set for all of the component SVMs in the ensemble. In practice, we use the set of features included in at least one out of every five models.

n	minimum AUC	mean AUC	maximum AUC
1	.5828	.6959	.7712
3	.6486	.7246	.7853
5	.6786	.7374	.7940
10	.6925	.7501	.7937

Table 1. Minimum, mean, and maximum 10-fold AUCs by ensemble size across 100 trials. The SVM feature set was composed of lateral tumor position, superior-inferior tumor position, performance status, and maximum dose to the heart.

4.4. Probabilistic Tuning

After the feature selection step, the output of each component SVM is tuned with an implementation of Lin et al.'s refinement of Platt's method (see Section 2.4) [12]. The decision function scores for input are generated by testing using a 10-fold cross-folding of the training set.

5. Results and discussion

We trained a series of 5 classifier ensembles using leave-one-out. The most commonly selected features across all the folds are the lateral position of the tumor (COMLAT), the superior-inferior position of the tumor (COMSI), the performance status of the patient (general health as evaluated by a physician), and the maximum dose to the heart. These features have all been identified as important RP factors in previous research [9, 7, 4]. Throughout this section, we will use this feature set as an approximation of the features set that would be selected by a sufficiently large collection of SVMs during feature selection within a fold.

To test for synergies arising from the ensemble method, we evaluated paired differences in 10-fold AUC for 100 different foldings using n = 1, 3, 5, 10 component SVMs. The outputs of the component SVMs were not tuned. Instead of repeatedly performing feature selection, the feature set containing COMSI, COMLAT, performance status, and maxi-

n	1	3	5	10
1	.2040	.0693	.0317	.7593
3		.4010	.2930	.6738
5			.3898	.6195
10				.6771

Table 2. Jarque-Bera test p-values for paired differences in AUC. Diagonal contains p-values for the individual sets.

mum dose to the heart was used. Feature scaling was still allowed during model building, however, via kernel σ selection. AUC summaries from the trials are shown in Table 1. These AUCs are not directly comparable to the prior SVM result in [2] because of patient population differences – patients in our data only received treatment for non-small-cell lung cancer. The seeming inconsistency with our prior result in [17] can be explained, in part, by (1) the difference in the number of folds (2) the uniform set of features across all component SVMs (3) differences in the partitions underlying the component SVMs.

To perform a paired Student's t-test to detect differences in mean model performance, the underlying distribution of differences must be approximately normal. Jarque-Bera tests reject normality at the 5% significance level only for the n=5 v. n=1 case (p-values are shown in Table 2) [10]. For the other pairs, a series of paired Student's t-test were performed with the hypotheses:

- $\mathrm{H}_{\mathrm{null}} : \mu_{X-Y} = 0$
- $\mathrm{H}_{\mathrm{alt}} : \mu_{X-Y} > 0$

, where X is the distribution of AUCs for larger classifier. The null hypothesis was rejected for all comparisons at the 5% significance level in favor of the one-tailed alternative (see Table 3). This suggests that larger ensembles outperform smaller ensembles and single classifiers for the selected sizes. Thus, synergy can be captured without introducing methodological differences in component classifiers as seen in [3]. It it important to note, however, that the assumption of independence between pairs had to be relaxed since all foldings contain the same underlying patient data.

Next, we consider ensembles with tuned output. Since patient outcomes are binary, the quality of probabilistic outputs cannot be directly measured. AUC, however, is still an important metric because it is based on the relative decision function scores. A low AUC for an ROC curve constructed from probability estimates implies poor relative probabilities.

Hope et al. evaluate model probability outputs graphically by binning patients by predicted risk and plotting the predicted and actual incidences of RP within each bin [9]. We do the same using LOO probability scores for ensembles with 20 component SVMs. The ROC curve, with AUC=0.7312, is shown in Fig. 3.

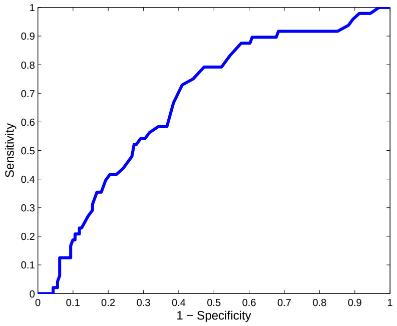

Figure 3. ROC built from LOO cross-validation scores for a n=20 SVM ensemble with probabilistic outputs.

n	3	5	10
1	3.9964e-10	*	7.6572e-29
3		2.9434e-06	1.1102e-16
5			1.4991e-07

Table 3. One-tailed Student t-test p-values for paired differences in AUC. * indicates normality assumption was violated.

Fig. 4 shows the predicted and actual RP incidence rates in 6 groups binned by predicted RP. The higher actual RP incidence rate in Bin 3 compared to Bin 4 is indicative of poor relative rankings. This discrepancy can be expected since the AUC of 0.7312 reflects a 27% probability that a random patient that does not develop RP will receive a higher predicted risk than a random patient that will develop RP. The over-estimation of RP risk in the lower bins can be explained by the averaging performed during model fusion. The lowest fused probability is 8.04%, while the lowest single SVM probability estimate is 1.26%.

Fig. 5 shows predicted and actual RP binned rates when predicted probabilities are calculated as the mean of 100 non-tuned binary-outcome SVMs – following the main idea in [3]. The large over-estimation of risk in Bin 5 and Bin 6 suggest that the mean binary-outcome is not a suitable proxy for RP risk probability.

While quality of absolute probability estimates for both methods is debatable, the ability to assign a patient to a relative risk group is useful in a clinical setting.

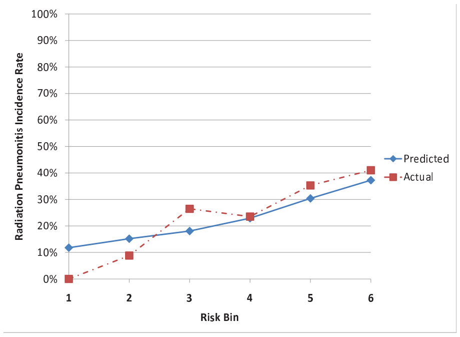

Figure 4. RP incidence probabilities binned by Platt-tuned predicted probability.

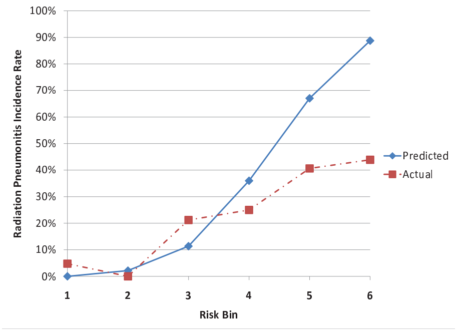

Figure 5. RP incidence probabilities binned by binary-averaged predicted probability.

6. Conclusion

We have presented a feature-ranking step for maintaining parsimony when modeling radiation pneumonitis with an ensemble of support vector machines. We then showed that larger ensembles produce improved 10-fold cross-validated AUCs at a statistically significant level. Finally, we demonstrated that generating probability estimates with Platt's method from the component SVMs provides benefits for clinical use. However, these potential benefits are limited by errors in relative risk assessments, as explained by the area under the receiver operating characteristic curve.

References

[1] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Mach. Learn.*, 46(1-3):131–159, 2002.

[2] S. Chen, S. Zhou, F.-F. Yin, L. B. Marks, and S. K. Das. Investigation of the support vector machine algorithm to predict lung radiation-induced pneumonitis. *Medical Physics*, 34(10):3808–3814, 2007.

[3] S. K. Das, S. Chen, J. O. Deasy, S. Zhou, F.-F. Yin, and L. B. Marks. Combining multiple models to generate consensus: Application to radiation-induced pneumonitis prediction. *Medical Physics*, 35(11):5098–5109, 2008.

[4] J. Deasy, M. Trovo, E. Huang, Y. Mu, I. E. Naqa, and J. Bradley. High-dose heart irradiation is a statistically significant risk factor for radiation pneumonitis within logistic-multivariate modeling. *International Journal of Radiation Oncology Biology Physics*, 72(1):S119–S119, 2008.

[5] C. Dehing-Oberije, D. D. Ruysscher, A. van Baardwijk, S. Yu, B. Rao, and P. Lambin. The importance of patient characteristics for the prediction of radiation-induced lung toxicity. *Radiotherapy Oncology*, 2009.

[6] T. Fawcett. An introduction to roc analysis. *Pattern Recogn. Lett.*, 27(8):861–874, 2006.

[7] M. V. Graham, J. A. Purdy, B. Emami, W. Harms, W. B. W, M. A. Lockett, and C. A. Perez. Clinical dose-volume histogram analysis for pneumonitis after 3d treatment for non-small cell lung cancer (nsclc). *Int J Radiat Oncol Biol Phys.*, 45(2):323–329, 1999.

[8] D. J. Hand and R. J. Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Mach. Learn.*, 45(2):171–186, 2001.

[9] A. J. Hope, P. E. Lindsay, I. E. Naqa, J. R. Alaly, M. Vicic, J. D. Bradley, and J. O. Deasy. Modeling radiation pneumonitis risk with clinical, dosimetric, and spatial parameters. *Int J Radiat Oncol Biol Phys.*, 65:112–124, 2006.

[10] C. M. Jarque and A. K. Bera. A test for normality of observations and regression residuals. *International Statistical Review*, 55(2):163172, 1987.

[11] S. S. Keerthi and C.-J. Lin. Asymptotic behaviors of support vector machines with gaussian kernel. *Neural Computation*, 15(7):1667–1689, 2003.

[12] H.-T. Lin, C.-J. Lin, and R. C. Weng. A note on platts probabilistic outputs for support vector machines. *Machine Learning*, 68(3):267–276, 2007.

[13] I. E. Naqa, J. D. Bradley, and J. O. Deasy. Nonlinear kernel-based approaches for predicting normal tissue toxicities. *Machine Learning and Applications, Fourth International Conference on*, 0:539–544, 2008.

[14] J. C. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines, 1998.

[15] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

[16] T. J. Robnett, M. Machtay, E. F. Vines, M. G. McKenna, K. M. Algazy, and W. G. McKenna. Factors predicting severe radiation pneumonitis in patients receiving definitive chemoradiation for lung cancer. *Int J Radiat Oncol Biol Phys*, 48(1):89–94, 2000.

[17] T. W. Schiller, Y. Chen, I. E. Naqa, and J. O. Deasy. Modeling radiation-induced lung injury risk with an ensemble of support vector machines. Submitted to Neurocomputing.

[18] V. Vapnik and O. Chapelle. Bounds on error expectation for support vector machines. *Neural Computation*, 12:2013–2036, 2000.