1-1-2009

Ensemble Support Vector Machine Models of Radiation-Induced Lung Injury Risk

Todd Schiller

WASHINGTON UNIVERSITY IN ST. LOUIS

School of Engineering and Applied Science

Department of Computer Science and Engineering

Thesis Examination Committee:
Yixin Chen, Chair
Ron Cytron
Joseph O. Deasy
Mark Franklin
Issam El Naqa

ENSEMBLE SUPPORT VECTOR MACHINE MODELS OF RADIATION-INDUCED LUNG INJURY RISK

by

Todd Wademan Schiller

A thesis presented to the School of Engineering
of Washington University in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

August 2009
Saint Louis, Missouri

ABSTRACT OF THE THESIS

Ensemble Support Vector Machine Models of Radiation-Induced Lung Injury Risk

by

Todd Wademan Schiller

Master of Science in Computer Science

Washington University in St. Louis, 2009

Research Advisor: Professor Yixin Chen

Patients undergoing radiation therapy can develop a potentially fatal inflammation of the lungs known as radiation pneumonitis (RP). In practice, modeling RP factors is difficult because existing data are under-sampled and imbalanced. Support vector machines (SVMs), a class of statistical learning methods that implicitly maps data into a higher dimensional space, is one machine learning method that recently has been applied to the RP problem with encouraging results. In this thesis, we present and evaluate an ensemble SVM method of modeling radiation pneumonitis. The method internalizes kernel/model parameter selection into model building and enables feature scaling via Olivier Chapelle's method. We show that the ensemble method provides statistically significant increases to the cross-folded area under the receiver operating characteristic curve while maintaining model parsimony. Finally, we extend our model with John C. Platt's method to support non-binary outcomes in order to augment clinical relevancy.

Acknowledgments

I would like to thank the individuals that have helped lead my exploration of science throughout my life: Dr. Sanjiv K. Bhatia, Dr. Justin Fay, Dr. Aaron Stump, Dr. Yixin Chen, Dr. Mark Franklin, Dr. Issam El Naqa, Dr. Joseph Deasy, and Dr. Ron Cytron. Simple words cannot express how appreciative I am of the knowledge with which you have entrusted me - I can only work to ensure that my present and future endeavors evince the greatness of this knowledge.

And to all the others that have played less official roles in my academic life: I express my deepest gratitude for lessons learnt, things discovered, and friends made.

Todd Wademan Schiller

Washington University in Saint Louis
August 2009

Dedicated to my parents.

Contents

List of Tables

List of Figures

Preface

This thesis is composed of two main chapters. In Chapter 2, we present an improved binary-outcome model for predicting radiation pneumonitis in patients undergoing radiation therapy. In Chapter 3, we adjust the model to support a more clinically relevant view of risk. Each part is meant to be able to stand alone as an innovative contribution to the field of patient outcome modeling. Such intention is drawn, in part, by the circumstances under which the chapters were researched and written: the second chapter was written as a submission for a special issue of *Neurocomputing* on subspace learning; later, the third chapter was written as a submission to a special session on modeling treatment outcomes in cancer and radiation therapy at the International Conference on Machine Learning and Applications.

Chapter 1

Introduction

Radiation pneumonitis (RP) is an inflammation of the lungs that presents within six months of thoracic radiation therapy. RP is potentially fatal, but symptoms can be as mild as a cough. Numerous factors, such as gender [14, 47], maximum dose [16], and tumor location [32] have been associated with radiation pneumonitis.

Accurate models of the risks stemming from patient irradiation allow clinicians to design effective radiation plans while controlling potential side effects. For RP, an ideal model would output the exact probability that a patient will develop clinically significant radiation pneumonitis. Such a model, however, does not exist yet – but not for a lack of trying. In fact, lung-injury prediction research has a rich history. Though a full review is beyond the scope of this thesis, a brief chronology helps to provide context for our work:

By the early 1970s, a growing set of factors had been identified as effecting RP risk [20, 34]. During the decade, RP research focused on describing the effects of various drugs on RP incidence. For example, in 1973, Wara et al. presented a probit model to evaluate the effect of dactinomycin administration on RP incidence [57].

Radiation pneumonitis research in the 1980s lacked a cohesive theme. Rothwell et al. showed that RP was strongly linked to irradiation volumes in breast cancer patients [48]. Koga et al. found that age was a significant factor effecting RP severity [36]. There was also a push to gain a better understanding of the biology underlying radiation pneumonitis [26, 56].

In the 1990s, efforts began to model patient outcomes for various conditions with machine learning techniques. For example, in 1997, Cooper et al. evaluated 8 statistical and machine learning models for predicting pneumonia mortality [13]. They found that an artificial neural network provided the best performance (though not necessarily statistically significant). In 1998, Munley et al. demonstrated that neural networks could also produce promising models of radiation-induced lung injury [43].

In the past ten years, many more machine learning techniques have been applied to the lung-injury prediction problem with varying levels of success; for example – self-organizing maps [11], decision trees [15], and support vector machines [10, 19]. However, an increase in available data also enabled a reexamination of more traditional statistical models [50, 32].

In this thesis, we build an improved model of RP with support vector machines (SVMs), a class of statistical learning methods. SVMs project their input into a higher-dimensional feature space in which the data is separable by a hyperplane. The mapping allows SVMs to capture complex relationships between features/factors. SVM-based models of RP have shown encouraging results [10, 14, 19].

The overriding purpose of this work is to improve the current state of SVM models of RP and to highlight issues affecting model quality. The primary vehicle used to achieve this purpose is an ensemble SVM method we present. The ensemble model combines the output from numerous SVMs to produce a higher-quality prediction function.

In Chapter 2, we formalize the ensemble SVM method and present results that suggest increased performance over previous SVM models. We explain the performance benefits by looking at the synergies captured by the model.

In Chapter 3, we adjust the feature selection method of our ensemble method in order to support model parsimony and statistically show that the ensemble method provides improved performance. Finally, we introduce a tuning step that allows the model to produce probabilistic risk estimates and discuss the step's positive impact on clinical relevance.

In Chapter 4, we offer concluding remarks and provide guidance for future research.

Chapter 2

Modeling Radiation-Induced Lung Injury Risk with an Ensemble of Support Vector Machines[1]

2.1 Introduction

Radiation Pneumonitis (RP) is a potentially fatal inflammation of the lungs that can occur as a result of thoracic radiation therapy (See Figure 2.1). Symptoms ranging from cough and fever to acute respiratory distress present themselves within six months of therapy. Because of the wide range of severity, institutions develop grading scales to characterize radiation pneumonitis events. Washington University's scale is shown in Table 2.1.

Numerous factors have been identified as contributing to radiation pneumonitis risk. Factors shown to be correlated with RP include treatment factors such as equivalent uniform dose [14, 10] and dose location [32, 60, 53] as well as clinical factors like

[1]Submitted on May 10, 2009 to the *Neurocomputing* special issue on subspace learning.

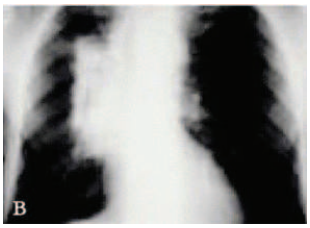

Figure 2.1: CT scan showing radiation-induced inflammation in the right lung (left in the picture) [37].

Table 2.1: Radiation pneumonitis grade definition from [32]

Washington University Lung Toxicity Criteria	
Grade	Definition
1	Mild symptoms of dry cough or dyspnea on exertion not requiring clinical intervention or radiographic evidence of pneumonitis without clinical symptoms
2	Steroids given for clinically significant pulmonary symptoms
3	Hospitalization for symptoms of dyspnea requiring supportive care (oxygen)
4	Severe respiratory insufficiency/continuous oxygen or assisted ventilation
5	Fatal

gender [14, 47]. Many of the factors individually correlated with RP are highly intercorrelated [32]. Therefore, attempts to construct parsimonious models of radiation pneumonitis typically argue for a small subset of factors. For example, Das et al. identify chemotherapy, equivalent uniform dose, gender, and squamous cell histology as significant [14].

Modeling radiation pneumonitis is a particularly challenging problem because existing data is under-sampled – the ratio of variable factors to the number of patients is large – and unbalanced. Recently, the academic and medical community has seen an increased interest in applying machine learning techniques to predicting radiation pneumonitis risk. In particular, support vector machines (SVMs), which have been successfully used in domains ranging from cancer classification [28, 23] to image retrieval [52, 61], are now being applied to the RP modeling problem with promising results [10, 19].

In this paper, we introduce three innovations for modeling binary RP risk with support vector machines: (1) Utilizing an ensemble of SVMs to address data imbalance and boost performance (2) Feature scaling during model building to complement forward feature selection (3) Performing parameter selection concurrently with model building. We show that our model outperforms previous SVM models by comparing the area under the cross-validated receiver operating characteristic curves (ROC).

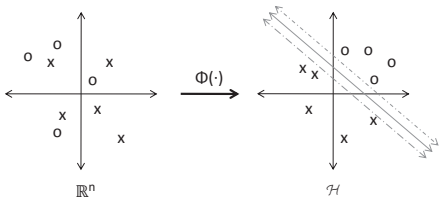

Figure 2.2: SVM classification. Left: Two classes of instances. Right: Instances in the implicit space, separated by the maximum-margin hyperplane; the dashed lines denote the margin.

In the next section, we provide a brief explanation of support vector machine classification. In Section 2.3, recent related literature and models are discussed. Then, in Section 2.4, we describe a novel SVM approach for modeling RP risk. In Section 2.5, we evaluate the model in relation to previous models. Finally, we offer concluding remarks in Section 2.6.

2.2 Background information

In this section, a brief background of classification methods is presented. The section first formalizes binary classification and support vector machine training. Feature selection and methods for model evaluation are then discussed.

2.2.1 Binary classification

The goal of binary classification is to construct a mapping function $f : \mathbb{R}^n \to \{-1, 1\}$ that maps an input vector to a label. In supervised learning, a set of input-label pairs,$\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_l, y_l)\}$, is used to train the classifier. The trained model should minimize model error when applied to future data.

2.2.2 Support vector machines

Support vector machines (SVMs) are a class of statistical learning methods that permit input data to be implicitly mapped into higher, possibly infinite, dimensional

spaces. Each potential mapping $\phi : \mathbb{R}^n \to \mathcal{H}$ produces a different SVM. Instead of explicitly mapping the input using ϕ, however, a kernel function $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defining the inner product in $\mathcal{H}$ implicitly maps the data. One popular kernel is the Gaussian radial basis function (RBF):

$$K_\sigma(x, y) = \exp\left(-\sum_i \frac{(x_i - y_i)^2}{2\sigma_i^2}\right) , \tag{2.1}$$

where σ is a vector of scaling factors.

The SVM training process finds the maximum-margin hyperplane separating the classes in the implicit space (Figure 2.2). Training results in a binary decision function of the form $f(x) = (w) \cdot \phi(\mathbf{x}) + b$. For separable data, the SVM training problem is the following optimization problem:

$$\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^T\mathbf{w} \tag{2.2}$$

subject to:

$$y_i(\mathbf{w}^T\phi(x_i) + b) \geq 1 \ .$$

Though the SVM can be trained using the primal (see [6] and [41]), the dual is typically solved instead:

$$\max_\alpha \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{2.3}$$

subject to:

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

The corresponding decision function is given by:

$$f(x) = \sum_i \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \ . \tag{2.4}$$

For non-separable datasets, a complexity constant can be introduced to permit training error. This is the class of *soft-margin* SVMs. Though the complexity parameter is often introduced into the model as a constraint on the Lagrangian multipliers in Equation 2.3, we instead choose to extend the kernel as in [8, 58]:

$$\mathbf{K} \leftarrow \mathbf{K} + \frac{1}{C}\mathbf{I} \ . \tag{2.5}$$

In practice, given a complexity parameter and a kernel, the SVM is trained using an algorithm such as sequential minimal optimization [45]. The proper complexity and kernel parameters are chosen by a naive enumeration over the parameter space, retraining the model each time. Chapelle et al., however, offer an alternative method for selecting parameters in which alternating SVM training and gradient descent steps are used to minimize the estimated generalization error [8].

2.2.3 Feature selection

As the number of features in the input increases relative to the number of significant features, models take longer to construct and also become less optimal (the curse of dimensionality). The goal of feature selection is to pick a subset of features such that the expected generalization error is minimized.

Let $\theta \in \{0,1\}^n$ be a feature selection vector providing a preprocessing of the data: $\mathbf{x} \rightarrow (\mathbf{x} * \theta)$ and $\tau : \{0,1\}^n \rightarrow \mathbb{R}$ be the expected generalization error when using preprocessing θ. The feature selection problem can then be expressed formally as [58]:

$$arg\, min_{\theta \in \{0,1\}^n} \tau(\theta) \ . \tag{2.6}$$

Since an exhaustive search of the 2^n possible subsets is generally intractable, other approaches are used.

2.2.4 Statistical model evaluation

Models are typically tested on a validation set, a set of data that is not used when constructing the model. When data is scarce, however, it is undesirable to exclude a subset of data from training. Therefore, cross-validation is used. In cross-validation, the available data is split into mutually exclusive subsets. Each subset is used as a validation set one time while the model is constructed using the remaining subsets. The results are then compiled to estimate the model's performance. When data is particularly scarce, leave-one-out (LOO) cross-validation is used. In LOO, each input-label pair $(\mathbf{x}_i, y_i)$ is used as a validation set exactly once while the model is trained using the other data.

Given a set of input-label pairs,$\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_l, y_l)\}$, the sensitivity and specificity of a binary classifier are:

$$\text{sensitivity} = \frac{TP}{TP + FN} \tag{2.7}$$

$$\text{specificity} = \frac{TN}{TN + FP} \ , \tag{2.8}$$

where TP, FP, TN and FN are the number of true positives, false positives, true negatives, and false negatives, respectively.

The receiver operating characteristic (ROC) curve is a plot of sensitivity against (1 - specificity) for varying decision function thresholds. The area under the ROC curve, the AUC, is used as a single-variable metric of model performance. An AUC of 0.5 corresponds to the performance of a random classifier. If the decision function scores are sorted in ascending order, the AUC can be estimated using:

$$\hat{A} = \frac{S_0 - n_0(n_0+1)/2}{n_0 n_1} \ , \tag{2.9}$$

where n_0 is the number of positive instances, n_1 is the number of negative instances, and S_0 is the rank sum of the positive instances [30].

Another single-value measure of model performance is the Matthews correlation coefficient (MCC):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{2.10}$$

An MCC of +1.0 corresponds to a perfect classifier, while an MCC of 0.0 corresponds to a random classifier.

2.3 Related work

Hope et al. construct a logistic regression model for radiation pneumonitis risk in patients undergoing radiation therapy for non-small-cell lung cancer. Features are selected via statistical bootstrapping. The resulting model is evaluated by first binning the instances according to predicted risk and then comparing the predicted and the actual RP incidence within the bin [32]. Gayou et al. instead use a genetic algorithm to select features for the logistic regression. The algorithm's fitness function is based on the model's predictive ability and on the statistical significance of the constituent features, the latter being included to prevent over-fitting. The choice of fitness function as a limiting factor of the model's actual performance is emphasized [24].

Chen et al. use a binary-outcome SVM model with an RBF kernel for predicting clinically significant RP events (Grade 2+ pneumonitis). The dataset was constructed from a study of 235 patients receiving three-dimensional conformal radiotherapy. Feature selection is performed based on improvement to the area under a cross-validated ROC curve. A model built from all variables is compared via ROC analysis to a

model with only dosimetric variables. For 10-fold cross-validated testing, the areas under the ROC curves are 0.71 for the dosimetric model and 0.76 for the full model [10]. Das et al. extend this work by including the SVM model in an ensemble of classifiers that include a feed-forward neural network [12], a decision tree [15], and a self-organizing map [11]. The cross-folded binary results of the classifiers are averaged to produce a real-valued risk estimate. An AUC of 0.79 is found for the combination of 100 cross-validated predictions from each of the models [14].

Using the same patient population, Dehing-Oberije et al. build uni- and multi- variate models with SVMs. Uni-variate models are built using V_{20} – the volume of the lung receiving at least 20 Gy – and the mean dose to the lung (MLD). The models are evaluated using LOO AUC. The highest AUC, 0.62, is achieved by the multi-variate model. The difference in AUC from [10] is attributed to differences in radiation doses [17].

El Naqa et al. also use SVMs to construct a binary model of RP risk using dosimetric and non-dose variables. The performance of features selected using logistic regression are compared to those chosen by recursive feature elimination (see [28]). The SVM built with features from the logistic model is shown to outperform those chosen by SVM-RFE – an MCC of 0.34 compared to 0.22. The model MCC of 0.34 constitutes a 46% improvement over the previous logistic model [19].

The idea of aggregating the output of classifiers trained on sampled data can be traced back to Breiman's work in 1996 [3]; Breiman's "bagging" method is now standard fare in data mining textbooks [29]. However, performance differences arising from implementation and domain variations warrant application specific studies.

For example, Tao et al. apply an ensemble SVM to the problem of image retrieval. Since the image retrieval domain also deals with unbalanced data, they employ a method similar to the one we present in Section 2.4.2 to produce balanced training data for the component classifiers. The difference is that their method selects negative instances via sampling with replacement while ours draws the negative instances from a random permutation. In addition, instead of performing feature selection, they build component classifiers with randomly sampled feature sets [51]. While this approach addresses the under-sampling problem, it is of limited use in domains (such as RP) where is useful to identify a core set of important features. Li et al. combine these

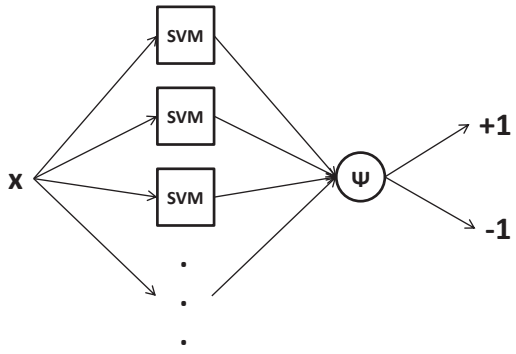

Figure 2.3: SVM ensemble. Decision function scores from each SVM are combined using fusion function Ψ.

methods with cotraining to better meet the relevance feedback paradigm common in image retrieval [39]. Other areas in which SVM ensembles have been applied include face detection [4] and cancer recognition [54].

Selecting training subsets in the presence of unbalanced data is a field in its own right [38, 9, 1, 2, 51]. Using the training subsets in an ensemble learner can provide many new challenges and opportunities. For instance, Hido and Kashima recently suggested under-sampling with a negative binomial distribution to produce roughly balanced subsets for training an ensemble. The method may be more robust than those that rely on equally balanced subsets [31].

2.4 Radiation pneumonitis risk model

In this section, we present the construction of our binary radiation pneumonitis model. The output of a collection of SVMs (Figure 2.3) is synthesized to produce a single decision function.

2.4.1 Data description

The dataset consists of 209 patients treated with radiation for non-small-cell lung cancer between 1991 and 2001. WUSTL Grade 2+ and RTOG Grade 3+ RP events were

considered significant for data labeling. Of the 209 patients, 48 (23%) exhibited clinically significant radiation pneumonitis events. The data include clinical, treatment, and location variables including, but not limited to: age, gender, performance status, smoking, treatment time, concurrent chemotherapy, and tumor-position. Some features, such as performance status – the general health of the patient – were determined by the patient's physician. Tumor position is recorded using a series of variables including lateral position (COMLAT), superior-inferior position (COMSI), and anterior-posterior position (COMAP). In addition, a series of dosimetric variables are also included in the data:

- D_X [heart, lung]: minimum dose to X% volume of the heart or lung, respectively
- V_X [heart, lung]: volume of the heart/lung receiving at least X Gy dose
- MOH_X [heart, lung]: mean of the hottest dose for X% of the heart/lung.

A Monte Carlo-based method was used to correct dose heterogeneity effect [16]. Features selected by the ensemble SVM model will be discussed in more detail in Section 2.5. We scale each feature to the range [0,1].

2.4.2 Ensemble classifier

Since only 23% of the patients developed significant RP, naively training a classifier on the full dataset results in a biased classifier – in the extreme case, the classifier will predict that no new instances will exhibit RP.

To address the issue of unbalanced data, we partitioned the data into a collection of balanced subsets. Each part consists of all the positive RP instances and an equal number of instances drawn from a random permutation of the negative instances (shown in Figure 2.4). See Algorithm 1.

A classifier is built for each subset of the data, as described in Sections 2.4.3 and 2.4.4. The decision function for the ensemble classifier is given by:

$$f(x) = \Psi\left(f_1(x), \cdots, f_C(x)\right) \ , \tag{2.11}$$

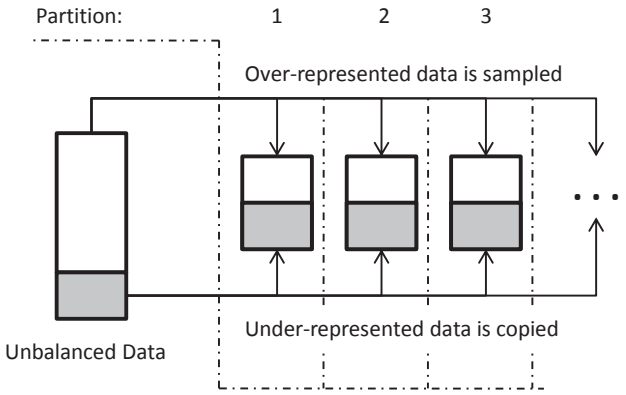

Figure 2.4: The data balancing process. The over-represented data is sampled according to a random permutation.

where $f_i(x)$ is the decision function for classifier i and $\Psi : \mathbb{R}^C \rightarrow \mathbb{R}$ is a fusion function. We calculate results for using both the mean and the median function for Ψ. The median is equivalent to a majority-vote when using an odd number of classifiers. It is possible to fuse the classifiers using a parametric scheme such as Adaboost [22], however, the theoretical and practical grounding for applying these methods to SVMs is still unclear [59, 40]. Therefore, we opt to use non-parametric fusion in this research.

2.4.3 SVM training and parameter selection

The model parameter C and RBF kernel width are not pre-selected. Instead, these parameters are selected at SVM training time using Chapelle et al.'s algorithm (a MATLAB implementation can be found at Olivier Chappelle's website)[8]. The algorithm alternates between SVM training and gradient descent steps to minimize expected generalization error. We use the algorithm to minimize the expected LOO error based on the span of the support vectors [7, 55]. The span S_p of support vector $\mathbf{x}_p$ is the minimum distance between $\phi(\mathbf{x}_p)$ and the set

$$\left\{ \sum_{i \neq p, \alpha_i^0} \lambda_i \phi(\mathbf{x}_i), \sum_{i \neq p} \lambda_i = 1 \right\} , \tag{2.12}$$

for $\sum \lambda_i = 1$ and $\alpha^{\mathbf{0}}$ are the values chosen by training the SVM in the dual.

Assuming the set of support vectors remains constant during LOO, the number of errors is:

$$T = \frac{1}{l} \sum_{p=1}^{l} \chi(\alpha_p^0 S_p^2 - 1) \ , \tag{2.13}$$

where l is the number of training instances, and

$$\chi(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases} \ .$$

We use the algorithm to select a scaling factor σ_i for each feature in the RBF kernel instead of selecting a single kernel width (see Equation 2.1).

2.4.4 Feature selection

Feature selection is performed for each classifier in the ensemble. As in [10], features are forward-selected by adding or substituting features that increase the 10-fold cross-validated AUC (on only the input-label pairs in the subset). New features are added or randomly substituted into the model until the AUC is no longer improved. The AUC is estimated using Equation 2.9. Forward selection is utilized for two reasons:

1. The features have previously been shown to be highly intercorrelated [32], making accurate backward selection difficult.

2. The existing body of literature suggests that RP can be modeled with relatively few features.

It should be noted that each time a model is built and evaluated for a subset of features, parameters C and σ are re-selected. This differs from previous work, in which final model and kernel parameters are selected prior to feature selection. We introduce an explicit cap for the number of features in an individual classifier in order to support the parsimony of the ensemble classifier.

Input: Positive instances, negative instances, number of partitions
Output: Balanced partitions
P = set of positive input-label pairs
$|P|$ = number of positive instances
$NegPerm$ = RandomPermutation(negative instances)
foreach *Partition* X **do**
    N = the next $|P|$ elements of $NegPerm$, re-permuting if necessary
    $X = P \cup N$
end

Algorithm 1: Creating balanced data partitions

2.5 Experimental results and discussion

Decision function scores for the ensemble are calculated using LOO cross-validation on the dataset. If an instance was used to build a particular classifier, that SVM is rebuilt without the instance (including reselecting model parameters C and σ). The scores are used to calculate the ROC curve and the AUC. Unlike during feature selection, the AUC is found via trapezoidal integration of the ROC. Models were created by using an ensemble of 3, 5, or 7 classifiers and by limiting each classifier to 3, 5, or 7 features. We will refer to the ensemble classifier with i classifiers and j maximum features as the i/j classifier. Five trials were performed for each ensemble classifier.

The mean fusion function outperformed the median function for 78% of the ensemble trials (with a mean difference to the AUC of 0.012). Therefore, we will only discuss classifiers using a mean to create fusion henceforth.

The min/mean/max results are shown for the */3 and */5 classifiers with a mean fusion function in Figure 2.5. The best mean AUC for a */3 classifier of 0.802 was obtained when 5 classifiers were used in the ensemble. The best for a */5 classifier, 0.818, was also obtained when 5 classifiers were used. For the */7 case (not shown), the best mean, 0.815, occurred when 7 classifiers were used.

We will use the mean 5/5 classifier results to evaluate our method in the context of previous work. The 5/5 model provides a better AUC mean and range when compared to the */3 class (see Figure 2.5). The 5/5 model uses more features, however, and

Table 2.2: Features selected by a 5/5 classifier with near-average performance. For classifiers with less than 5 features, the cross-validated AUC could not be increased by a round of substitution or addition of another feature. Scaling factors are shown in parenthesis. The corresponding ROC curves are shown in Figure 2.6.

Features Selected by an Average 5/5 Model		
1	COS Heart Z (.5815) D_{80} Lung MC (.1705)	Performance Status (.2597) COMLAT (.0726)
2	MOH_{60} Lung MC (.4783) COMSI (.2465)	COMAP (.2806) Performance Status (.2445)
3	Performance Status (.2815) MOH_{95} Lung MC (.1588) D_5 Lung MC (.1361)	MOH_5 Heart MC (.2147) D_{45} Lung MC (.1456)
4	MOH_{10} Heart MC (.3935) Performance Status (.1906)	D_{75} Lung MC (.3549)
5	D_{45} Lung MC (.3476)	MOH_5 Heart MC (.2728)

thus may be less parsimonious. Compared to the */7 class, the 5/5 results in a larger AUC while also using fewer features.

The features chosen by a nearly average 5/5 classifier (AUC=0.814) are shown in Table 2.2. This set of selected features includes tumor location features (COMLAT, COMSI, COMAP), performance status, and dosimetric parameters (D_X for heart and lung, MOH_X for heart and lung). As the dosimetric variables – D_X in particular – have previously been shown to be intercorrelated [32], it may be possible to further condense the feature space without significantly harming model performance.

The 5/5 ensemble classifier for binary RP prediction compares favorably to the work by Chen et. al that finds an AUC of 0.76. The results are not directly comparable, however, for two reasons: (1) we calculated the AUC using LOO whereas Chen et. al use 10-fold cross-validation; (2) our dataset is restricted to patients undergoing treatment for non-small-cell lung cancer as opposed to general lung cancer patients.

It should be noted that the component classifiers in this work typically underperform the resulting single SVM classifier in Chen et al.'s work. This can be explained by the data partitioning process in which only 28.2% of the RP-negative instances are included as training data for each classifier. Though the partitioning limits the performance of a single classifier, we believe it is important in the creation of synergies

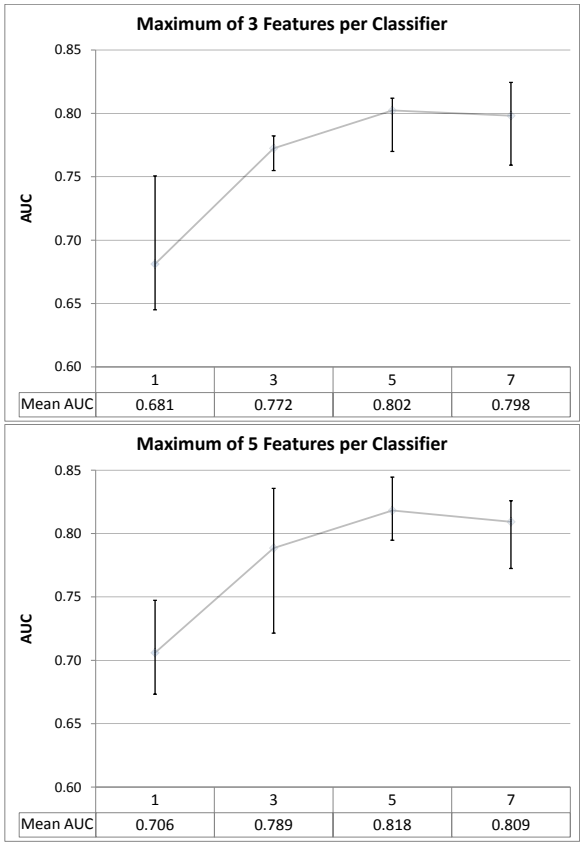

Figure 2.5: The mean AUC (across 5 trials) vs. the number of classifiers in the ensemble. The end points of the vertical bars denote the maximum and minimum AUCs. Top: Each classifier is limited to 3 features. Bottom: Each classifier is limited to 5 features.

during model fusion (model biases complement each other). Figure 2.6 shows the ROC of the near-average 5/5 model and its component classifiers.

This type of synergy is also described in Das et al.'s work on combining multiple classification methods for predicting RP [14]. Using 100 cross-validated predictions from each collection of classifier (an SVM, an NN, an SOM, and a decision tree) results in an AUC of 0.79. As with Chen et al.'s work in [10], the results aren't directly comparable since the patient populations and the method of calculating AUC differ. But, a couple insights can sill be made: (1) our model produces a similar performance using only a single type of classifier (2) ensemble/fusion classification is a promising way to take advantage of classifier bias.

The average 5/5 classifier also outperforms El Naqa et al.'s classifier in [19]. The LOO Matthews correlation coefficient in El Naqa's work is 0.34. The 5/5 classifier ensemble has a mean LOO MCC of 0.497 across the five trials. It should be noted, however, that the dataset used by El Naqa et al. does not include dosimetric variables for the heart. Therefore, for comparison, we tested the 5/5 classifier on the same data used in El Naqa's work. Across 10 LOO trials, an average MCC of 0.37 was obtained. For both data sets, the decision function threshold can be tweaked to obtain yet a higher MCC. By transitivity, the ensemble also compares favorably to the model in [32], which El Naqa's method outperforms by 46% (measured using MCC).

To investigate the role that the balanced partitioning scheme plays in model performance, we tested the performance of a 5/5 classifier with training subsets randomly drawn from the complete dataset with replacement. Across 5 trials, the mean LOO AUC is 0.73 (with a minimum and maximum of 0.69 and 0.77, respectively). The mean MCC was 0.20. The inferior AUC and MCC suggest that data balancing is an integral part of the presented ensemble method.

Parameter selection during model building is not free – the average feature selection time for a component classifier with a maximum of 3, 5, and 7 features is 36.0, 57.6, and 45.8 minutes respectively (across 100 trials on Intel Core 2 Q6600 2.4 GHz machines with 2GB memory). The seemingly anomalous */5 and */7 running times result from the maximum feature constraint being not binding for all SVMs. For comparison, the standard grid search + LIBSVM [5] approach takes approximately a minute for component feature selection (for a maximum of 3, 5, and 7 features). The increased running times are still practical, however, since: (1) feature selection for component classifiers is trivially parallelizable and (2) the training time is short relative to the length of potential clinical use.

Overall, the method performs favorably when compared to previous SVM methods. Using the same base feature selection methodology as in [10], creating an ensemble of SVMs, using gradient selection to perform parameter selection, and permitting each feature to be scaled individually has resulted in a performance increase. Though it is clear that model fusion is beneficial, the individual effects of the gradient selection and feature scaling are not clear. It will be important to isolate these effects in the

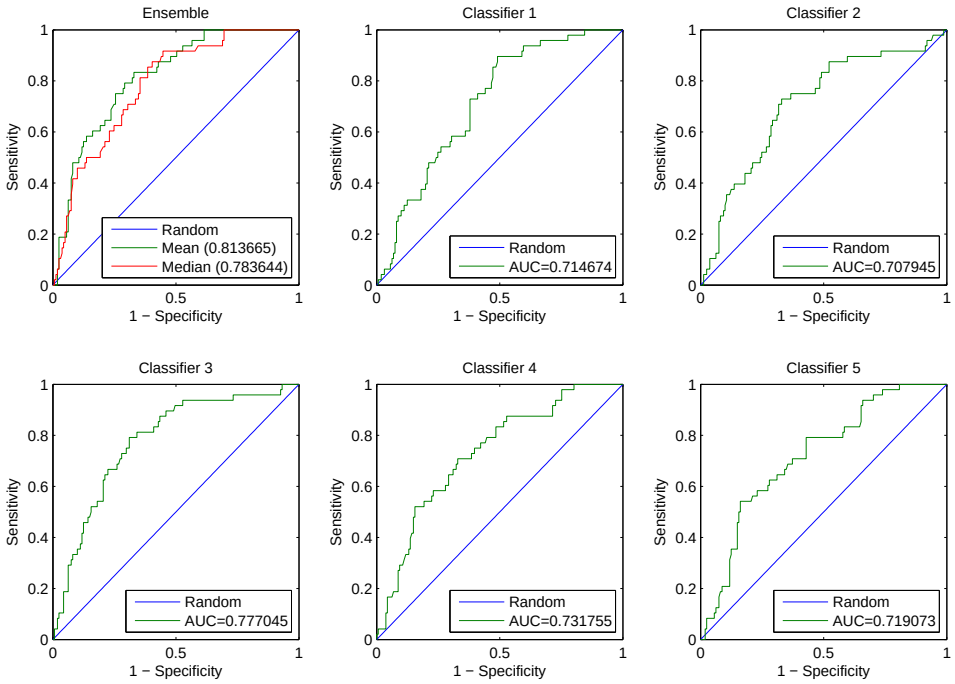

Figure 2.6: An average performance ensemble classifier with 5 component SVMs each restricted to 5 (the 5/5 model). The relatively weak classifiers complement each other to produce a strong ensemble classifier. The features used by each classifier are shown in Table 2.2.

future. It would also be interesting to see the effects of using our improved SVM model as part of a multi-classifier ensemble, such as the one presented in [14].

2.6 Conclusion

We have presented an SVM model of binary radiation pneumonitis risk with 3 innovations over previous models:

1. Utilizing an ensemble of SVMs to address data imbalance and to boost performance
2. Feature scaling during model building to complement forward feature selection
3. Performing parameter selection concurrently with model building

Using our methodology, we produced a set of models with varying numbers of classifiers and a maximum number of features per classifier. From these models, the ensemble with 5 component classifiers, with a maximum of 5 features each, is selected with an average leave-one-out AUC of 0.818. We showed that the average model of this type outperforms previous SVM and logistic models.

Chapter 3

Improving Clinical Relevance in Ensemble Support Vector Machine Models of Radiation Pneumonitis Risk[2]

3.1 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 [32] and gender [47, 14], 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 [14]. 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 [17] 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. 3.1).

[2]Submitted on August 1, 2009 to the *The Eighth International Conference on Machine Learning and Applications* (ICMLA 2009) special session on *Machine Learning Methods for Modeling Treatment Outcomes in Cancer*.

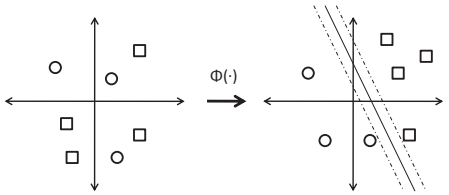

Figure 3.1: SVM classification: two classes of instances are mapped to an implicit space in which they are separable.

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 [46, 42].

In [49], 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 improve 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 3.3, we survey related work. In Section 3.4, we outline our improved ensemble SVM methodology. Results are presented and discussed in Section 3.5. Finally, we offer concluding remarks in Section 3.6.

3.2 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.

3.2.1 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):

$$K_\sigma(x, y) = \exp\left(-\sum_i \frac{(x_i - y_i)^2}{2\sigma_i^2}\right) , \tag{3.1}$$

where σ is a vector of scaling factors.

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

$$\max_\alpha \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{3.2}$$

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:

$$f(x) = \sum_i \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \ . \tag{3.3}$$

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:

$$\mathbf{K} \leftarrow \mathbf{K} + \frac{1}{C}\mathbf{I} \ , \tag{3.4}$$

where $\mathbf{I}$ is the identity matrix [8].

Kernel parameter σ and model parameter C are often selected prior to model building using grid-search [35].The optimization problem in Equation 3.2 can then be solved using Platt's sequential minimal optimization (SMO) method [45]. 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 [8].

3.2.2 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.

3.2.3 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 [21]. 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:

$$\hat{A} = \frac{S_0 - n_0(n_0+1)/2}{n_0 n_1} \ , \tag{3.5}$$

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 [30].

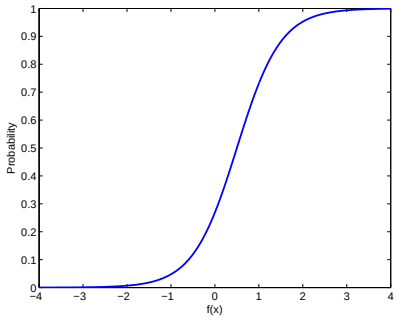

Figure 3.2: Sigmoid probability curve with A=-2 and B=1

3.2.4 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

$$P(y = 1|f) = \frac{1}{1 + \exp(Af + B)} \tag{3.6}$$

to the SVM output [46]. See Fig. 3.2 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:

$$\begin{aligned} t_+ &= \frac{N_+ + 1}{N_+ + 2} \\ t_- &= \frac{1}{N_- + 2} \ . \end{aligned} \tag{3.7}$$

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

$$\min_{A,B} \sum_i t_i \log(p_i) + (1 - t_i) \log(1 - p_i) \ , \tag{3.8}$$

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 [42].

3.3 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 [10]. 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 [19].

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 [10] by including the binary SVM model in a model that includes a feed-forward neural network, a decision tree, and a self-organizing map [14]. 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 [32].

3.4 Methods

This section briefly outlines our ensemble method in [49] and provides implementation details for the methods specific to this work. All the methods were implemented in Matlab 7.8.0 (R2009a).

3.4.1 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 [32]). 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 [16].

3.4.2 Ensemble of support vector machines

Instead of the single SVM approach used by Chen et al. [10], we use an ensemble of SVMs to address data imbalance and exploit potential synergies [49, 14]. As in our previous work in [49], 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 3.2.1. The method is used to minimize a support vector span estimate of the LOO error [55]. 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.

3.4.3 SVM feature selection

Features are selected according to a modified version of the AUC-maximizing forward selection algorithm in [10]. 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.

3.4.4 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 3.2.3) [42].

Table 3.1: 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.

n	minimum AUC	mean AUC	maximum AUC
1	0.5828	0.6959	0.7712
3	0.6486	0.7246	0.7853
5	0.6786	0.7374	0.7940
10	0.6925	0.7501	0.7937

The decision function scores for input are generated by testing using a 10-fold cross-folding of the training set.

3.5 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 [32, 25, 16]. 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 maximum 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 3.1. These AUCs are not directly comparable to the prior SVM result in [10] 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 [49] can be explained, in part, by (1) the difference in

Table 3.2: Jarque-Bera test p-values for paired differences in AUC. Diagonal contains p-values for the individual sets.

n	1	3	5	10
1	0.2040	0.0693	0.0317	0.7593
3		0.4010	0.2930	0.6738
5			0.3898	0.6195
10				0.6771

Table 3.3: One-tailed Student t-test p-values for paired differences in AUC. * indicates normality assumption was violated.

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

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 3.2) [33]. 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 3.3). 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 [14]. 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.

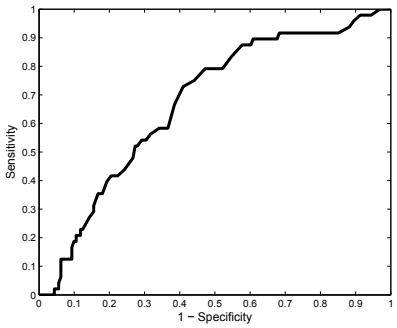

Figure 3.3: ROC built from LOO cross-validation scores for a n=20 SVM ensemble with probabilistic outputs.

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 [32]. 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. 3.3.

Fig. 3.4 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. 3.5 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 [14]. 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.

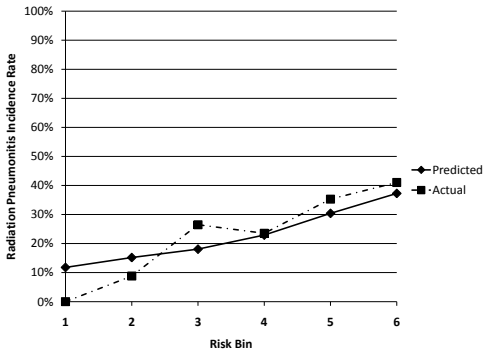

Figure 3.4: RP incidence probabilities binned by Platt-tuned predicted probability.

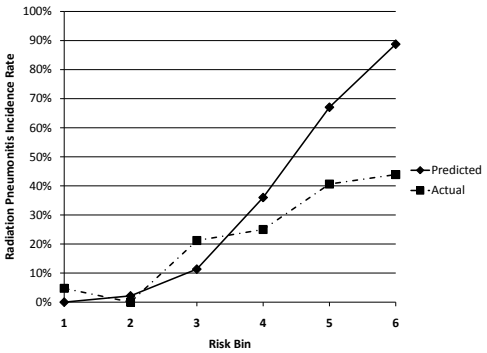

Figure 3.5: RP incidence probabilities binned by binary-averaged predicted probability.

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

3.6 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.

Chapter 4

Conclusion and Directions for Future Work

We have presented an ensemble SVM model of radiation pneumonitis that combines the strengths of individual SVM classifiers. Taking advantage of advances in general SVM theory, the model offers increased performance and probabilistic risk estimates while maintaining model parsimony.

Moving forward, there are many topics that should be investigated in SVM RP modeling. In particular, efforts to increase model AUC ought to continue. Models with higher AUCs not only serve as better binary-outcome models of RP risk, but also may provide more informative probability estimates.

Focus should also be given to improving the clinical relevance of AUC results. One potential improvement would be to restrict the AUC calculation to relevant/acceptable levels of specificity. Though estimating this partial AUC is less straight-forward, the body of theory is at the point where good estimates are possible [18].

Restricting the set of patients for which binary outcomes are predicted may also be advantageous. Allowing certain patients to be labeled as "hard-to-classify" by the model could result in an improved model for classifiable patients. Care must be taken, however, to ensure that the usefulness of the model is not undermined by excluding too many patients.

For any metric, the greatest future gains in model performance are most likely to come from the application of domain knowledge in data preprocessing – for example, Chen et al.'s work with equivalent uniform dose [10]. These methods help to capture complex and meaningful factor interactions that even the SVM kernel cannot.

The modeling of radiation pneumonitis risk with SVMs is still a new field. Our ensemble method provides a firm grounding for future research to maximize the performance of SVM-based models.

References

[1] N. Abe. Sampling approaches to learning from imbalanced datasets: active learning, cost sensitive learning and beyond. In *ICML-KDD Workshop: Learning from Imbalanced Data Sets*, 2003.

[2] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.*, 6:20–29, 2004.

[3] L. Breiman. Bagging predictors. *Mach. Learn.*, pages 123–140, 1996.

[4] L. Buciu, C. Kotropoulos, and I. Pitas. Combining support vector machines for accurate face detection. In *Proc. Intl. Conf. on Image Proc.*, pages 1054–1057, 2001.

[5] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. software available at http://www.csie.ntu.edu.tw/c̃jlin/libsvm, 2001.

[6] O. Chapelle. Training a support vector machine in the primal. *Neural Computation*, 19:1155–1178, 2007.

[7] O. Chapelle and V. Vapnik. Model selection for support vector machines. *Advances in Neural Info. Proc. Systems*, pages 230–236, 1999.

[8] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1):131–159, January 2002.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[10] 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, October 2007.

[11] S. Chen, S. Zhou, F. F. Yin, L. B. Marks, and S. K. Das. Using patient data similarities to predict radiation pneumonitis via a self-organizing map. *Physics in Medicine and Biology*, 53(1):203–216, January 2008.

[12] S. Chen, S. Zhou, J. Zhang, F. F. Yin, L. B. Marks, and S. K. Das. A neural network model to predict lung radiation-induced pneumonitis. *Medical physics*, 34(9):3420–3427, September 2007.

[13] G. F. Cooper, C. F. Aliferis, R. Ambrosino, J. Aronis, B. G. Buchanan, R. Caruana, M. J. Fine, C. Glymour, G. Gordon, B. H. Hanusa, J.E. Janosky, C. Meek, T. Mitchell, T. Richardson, and P. Spirtes. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artif. Intell. Med.*, 9(2):107–138, 1997.

[14] 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.

[15] S. K. Das, S. Zhou, J. Zhang, F. F. Yin, M. W. Dewhirst, and L. B. Marks. Predicting lung radiotherapy-induced pneumonitis using a model combining parametric lyman probit with nonparametric decision trees. *Int. J. Radiat. Oncol. Biol. Phys.*, 68(4):1212–1221, July 2007.

[16] J. O. Deasy, M. Trovo, E. X. Huang, Y. Mu, I. El Naqa, and J. D. Bradley. High-dose heart irradiation is a statistically significant risk factor for radiation pneumonitis within logistic-multivariate modeling. *Int. J. Radiat. Oncol. Biol. Phys.*, 72:S119, 2008.

[17] C. Dehing-Oberijea, D. De Ruysschera, A. van Baardwijka, S. Yub, B. Raob, and P. Lambina. The importance of patient characteristics for the prediction of radiation-induced lung toxicity. *Radiotherapy and Oncology*, 2008.

[18] L. E Dodd and M. S. Pepe. Partial auc estimation and regression. *Biometrics*, 59(3):614–623, 2003.

[19] I. El Naqa, J. D. Bradley, and J. O. Deasy. Non-linear kernel-based approaches for predicting normal tissue toxicities. In *Seventh Intl. Conf. on Machine Learning and Applications*, pages 539–544, 2008.

[20] J. C. Evans. Time-dose relationships of radiation; fibrosis of the lung. *Radiology*, 74, 1960.

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

[22] Y. Freund and E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Comp. and Sys. Sci.*, 55:119–139, 1997.

[23] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, October 2000.

[24] O. Gayou, S. K. Das, S. M. Zhou, L. B. Marks, D. S. Parda, and M. Miften. A genetic algorithm for variable selection in logistic regression analysis of radiotherapy treatment outcomes. *Medical Physics*, 35(12):5426–5433, 2008.

[25] M. V. Graham, J. A. Purdy, B. Emami, W. Harms, W. Bosch 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.

[26] N. J. Gross. The pathogenesis of radiation-induced lung damage. *Lung*, 159(1):115–125, 1981.

[27] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[28] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.

[29] J. Han and M. Kamber. *Data Mining: Concepts and Technique*. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, 2 edition, 2006.

[30] D. J. Hand and R. J. Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine Learning*, 45:171–186, 2001.

[31] S. Hido and H. Kashima. Roughly balanced bagging for imbalanced data. In *Proc. SIAM Intl. Conf. on Data Mining*, pages 143–152, 2008.

[32] A. J. Hope, P. E. Lindsay, I. El 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(1):112–124, May 2006.

[33] C. M. Jarque and A. K. Bera. A test for normality of observations and regression residuals. *International Statistical Review*, 55(2):163–172, 1987.

[34] F. L. Jennings and A. Arden. Development of radiation pneumonitis, time and dose factors. *Arch. Pathol*, 74:351–360, 1962.

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

[36] K. Koga, S. Kusumoto, K. Watanabe, K. Nishikawa, K. Harada, and H. Ebihara. Age factor relevant to the development of radiation pneumonitis in radiotherapy of lung cancer. *Int. J. Radiat. Oncol. Biol. Phys.*, 14(2):367–371, 1988.

[37] F.M. Kong, R. T. Hakena, A. Eisbrucha, and T. S. Lawrence. Non-small cell lung cancer therapy-related pulmonary toxicity: an update on radiation pneumonitis and fibrosis. *Seminars in Oncology*, pages S42–54, 2005.

[38] M. Kubat and S. Matwin. Addressing the curse of imbalanced training sets: one-sided selection. In *Proc. Fourteenth Intl. Conf. on Mach. Learn.*, pages 179–186, 1997.

[39] J. Li, N. M. Allinson, D. Tao, and X. Li. Multitraining support vector machine for image retrieval. *IEEE Transactions on Image Processing*, 15:3597–3601, 2006.

[40] X. Li, L. Wang, and E. Sung. Adaboost with svm-based component classifiers. *Engineering Applications of Artificial Intelligence*, 21:785–795, 2008.

[41] Z. Liang and Y. Li. Incremental support vector machine learning in the primal and applications. *Neurocomputing*, 72:2249–2258, 2009.

[42] H.-T. Lin, C.-J. Lin, and R. C. Weng. A note on platt's probabilistic outputs for support vector machines. *Machine Learning*, 68(3):267–276, 2007.

[43] M. T. Munley, J. Y. Lo, G. S. Sibley, G. C. Bentel, M. S. Anscher, and L. B. Marks. A neural network to predict symptomatic lung injury. *Physics in Medicine and Biology*, 44(9):2241–2249, 1999.

[44] T. L. Phillips, M. D. Wharam, and L. W. Margolis. Modification of radiation injury to normal tissues by chemotherapeutic agents. *Cancer*, 35(6):1678–1684, 1975.

[45] J. C. Platt. *Fast training of support vector machines using sequential minimal optimization*, pages 185–208. MIT Press, Cambridge, MA, USA, 1999.

[46] 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.

[47] 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, August 2000.

[48] R. I. Rothwell, S. A. Kelly, and C. A. Joslin. Radiation pneumonitis in patients treated for breast cancer. *Radiother. Oncol.*, 4(1):9–14, 1985.

[49] T. W. Schiller, Y. Chen, I. El Naqa, and J. O. Deasy. Modeling radiation-induced lung injury risk with an ensemble of support vector machines. 2009. Submitted to *Neurocomputing*.

[50] Y. Seppenwoolde, J. Lebesque, K. de Jaeger, J. Belderbos, L. Boersma, C. Schilstra, G. Henning, J. Hayman, M. Martel, and R. Ten Haken. Comparing different ntcp models that predict the incidence of radiation pneumonitis. *Int. J. Radiat. Oncol. Biol. Phys.*, 55(3):724–735, 2003.

[51] D. Tao, X. Tang, X. Li, and X. Wu. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28:1088–1099, 2006.

[52] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *Proceedings of the Ninth ACM Intl. Conf. on Multimedia*, pages 107–118, New York, NY, USA, 2001. ACM Press.

[53] K. Tsujino, S. Hirota, M. Endo, K. Obayashi, Kotani Y., M. Satouchi, T. Kado, and Takada Y. Predictive value of dose-volume histogram parameters for predicting radiation pneumonitis after concurrent chemoradiation for lung cancer. *Int. J. Radiat. Oncol. Biol. Phys.*, 55:110–115, 2003.

[54] G. Valentini, M. Muselli, and F. Ruffino. Cancer recognition with bagged ensembles of support vector machines. *Neurocomputing*, 56:461–466, 2004.

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

[56] J. A. Vergara and U. Raymond L. A. Thet. Changes in lung morphology and cell number in radiation pneumonitis and fibrosis: a quantitative ultrastructural study. *Int. J. Radiat. Oncol. Biol. Phys.*, 5:723–732, 1987.

[57] William M. Wara, Theodore L. Phillips, Lawrence W. Margolis, and Vernon Smith. Radiation pneumonitis: a new approach to the derivation of time-dose factors. *Cancer*, 32(3):547–552, 1973.

[58] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for svms. In *Advances in Neural Information Processing Systems*, volume 13, pages 668–674, 2000.

[59] J. Wickramaratna, S. Holden, and B. Buxton. Performance degradation in boosting. In *Proc. Second Intl. Workshop on Multiple Classifier Systems*, pages 11–21, 2001.

[60] M. Yamada, S. Kudoh, K. Hirata, T. Nakajima, and J. Yoshikawa. Risk factors of pneumonitis following chemoradiotherapy for lung cancer. *European Journal of Cancer*, 34(1):71–75, January 1998.

[61] L. Zhang, F. Lin, and Zhang B. Support vector machine learning for image retrieval. In *Proc. ICIP '01*, pages 721–724, 2001.

Vita

Todd Wademan Schiller

Date of Birth August 7, 1987

Place of Birth Creve Coeur, Missouri

Degrees B.S. Cum Laude, Computer Science, May 2009
M.S. Computer Science, August 2009

Societies Theta Xi Fraternity
Sigma Xi: The Scientific Research Society
Association for Computing Machines

Publications Stump, A., Deters, M., Petcher, A., Schiller, T., and Simpson, T. 2008. Verified programming in Guru. In *Proceedings of the 3rd Workshop on Programming Languages Meets Program Verification* (Savannah, GA, USA, January 20, 2009). PLPV '09. ACM, New York, NY, 49-58.

August 2009

Modeling Radiation Pneumonitis Risk, Schiller, M.S. 2009