Modeling radiation-induced lung injury risk with an ensemble of support vector machines

Todd W. Schiller, Yixin Chen

Department of Computer Science and Engineering, Washington University, St. Louis, Missouri

Issam El Naqa, Joseph O. Deasy

Department of Radiation Oncology, Washington University School of Medicine, Siteman Cancer Center, St. Louis, Missouri

Abstract

Radiation-induced lung injury, radiation pneumonitis (RP), is a potentially fatal side-effect of thoracic radiation therapy. In this work, using an ensemble of support vector machines (SVMs), we build a binary RP risk model from clinical and dosimetric parameters. Patient/treatment data is partitioned into balanced subsets to prevent model bias. Forward feature selection, maximizing the area under the curve (AUC) for a cross-validated receiver operating characteristic (ROC) curve, is performed on each subset. Model parameter selection and construction occurs concurrently via alternating SVM and gradient descent steps to minimize estimated generalization error. We show that an ensemble classifier with a mean fusion function, 5 component SVMs, and limit of 5 features per classifier exhibits a mean AUC of 0.818 – an improvement over previous SVM models of RP risk.

Key words: support vector machine, radiation pneumonitis, feature selection, ensemble learning, unbalanced data

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 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 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 [13, 10] and dose location [24, 40, 35] as well as clinical factors like gender [13, 32]. Many of the factors individually correlated with RP are highly intercorrelated [24]. 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 [13].

Figure 1: CT scan showing radiation-induced inflammation in the right lung (left in the picture) [25].

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 [20, 18] to image retrieval [34, 41], are now being applied to the RP modeling problem with promising results [10, 30].

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).

In the next section, we provide a brief explanation of support vector machine classification. In Section 3, recent related literature and models are discussed. Then, in Section 4, we describe a novel SVM approach for modeling RP risk. In Section 5, we evaluate the model in relation to previous models. Finally, we offer concluding remarks in Section 6.

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

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

Table 1: Radiation pneumonitis grade definition from [24]

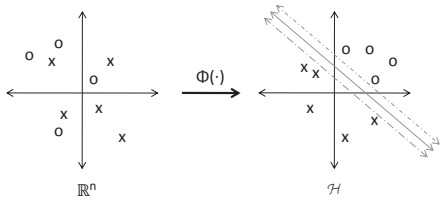

Figure 2: SVM classification. Left: Two classes of instances. Right: Instances in the implicit space, separated by the maximum-margin hyperplane; the dashed lines denote the margin.

and methods for model evaluation are then discussed.

2.1. Binary classification

The goal of binary classification is to construct a mapping function $f : \mathbb{R}^n \rightarrow \{-1, 1\}$ 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. 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 $\phi : \mathbb{R}^n \rightarrow \mathcal{H}$ produces a different SVM. Instead of explicitly mapping the input using ϕ, however, a kernel function $K : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ defining the inner product in $\mathcal{H}$ implicitly maps the data. One popular kernel is the Gaussian radial basis function (RBF):

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

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). 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:

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

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 [29]), the dual is typically solved instead:

$$\max_\alpha \sum_i \alpha_i - \frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j K(x_i, x_j) \tag{3}$$

subject to:

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

The corresponding decision function is given by:

$$f(x) = \sum_i \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b . \tag{4}$$

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 3, we instead choose to extend the kernel as in [8, 38]:

$$\mathbf{K} \leftarrow \mathbf{K} + \frac{1}{C}\mathbf{I} . \tag{5}$$

In practice, given a complexity parameter and a kernel, the SVM is trained using an algorithm such as sequential minimal optimization [31]. 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].

2.3. 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 [38]:

$$arg\,min_{\theta\in\{0,1\}^n}\tau(\theta)\ . \tag{6}$$

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

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:

$$\text{sensitivity} = \frac{TP}{TP + FN} \tag{7}$$

$$\text{specificity} = \frac{TN}{TN + FP}\ , \tag{8}$$

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:

$$\hat{A} = \frac{S_0 - n_0(n_0 + 1)/2}{n_0 n_1}\ , \tag{9}$$

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 [22].

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

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{10}$$

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

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 [24]. 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 [19].

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 [14], 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 [13].

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 [16].

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 [20]). 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 [30].

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 [21]. 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 4.2 to produce balanced training data for the component classifiers. The difference is that their method selects negative instances via sampling with re-

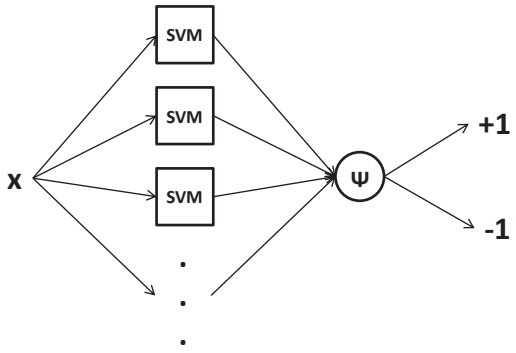

Figure 3: SVM ensemble. Decision function scores from each SVM are combined using fusion function Ψ.

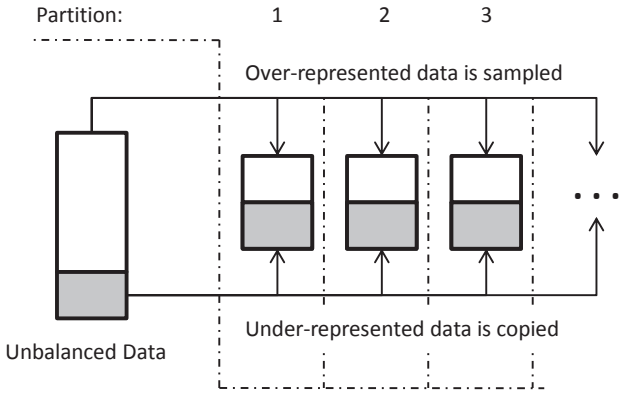

Figure 4: The data balancing process. The over-represented data is sampled according to a random permutation.

placement 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 [33]. While this approach addresses the undersampling 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 methods with cotraining to better meet the relevance feedback paradigm common in image retrieval [27]. Other areas in which SVM ensembles have been applied include face detection [4] and cancer recognition [36].

Selecting training subsets in the presence of unbalanced data is a field in its own right [26, 9, 1, 2, 33]. 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 [23].

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 3) is synthesized to produce a single decision function.

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 [15]. Features selected by the ensemble SVM model will be discussed in more detail in Section 5. We scale each feature to the range [0,1].

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 4). See Algorithm 1.

A classifier is built for each subset of the data, as described in Sections 4.3 and 4.4. The decision function for the ensemble classifier is given by:

$$f(x) = \Psi\left(f_1(x), \cdots, f_C(x)\right) , \tag{11}$$

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 [17], however, the theoretical and practical grounding for applying these methods to SVMs is still unclear [39, 28]. Therefore, we opt to use non-parametric fusion in this research.

4.3. SVM training and parameter selection

The model parameter C and RBF kernel width are not preselected. 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, 37]. The span S_p of support vector $\mathbf{x}_p$ is the minimum distance between $\phi(\mathbf{x}_p)$ and the set

$$\left\{ \sum_{i \neq p, \alpha_i^0} \lambda_i \phi(\mathbf{x}_i), \sum_{i \neq p} \lambda_i = 1 \right\}, \tag{12}$$

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:

$$T = \frac{1}{l} \sum_{p=1}^{l} \chi(\alpha_p^0 S_p^2 - 1), \tag{13}$$

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 1).

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 9. Forward selection is utilized for two reasons:

1. The features have previously been shown to be highly intercorrelated [24], 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

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 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 5). The 5/5 model uses more features, however, and 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. 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 [24], 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

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)

Table 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 6.

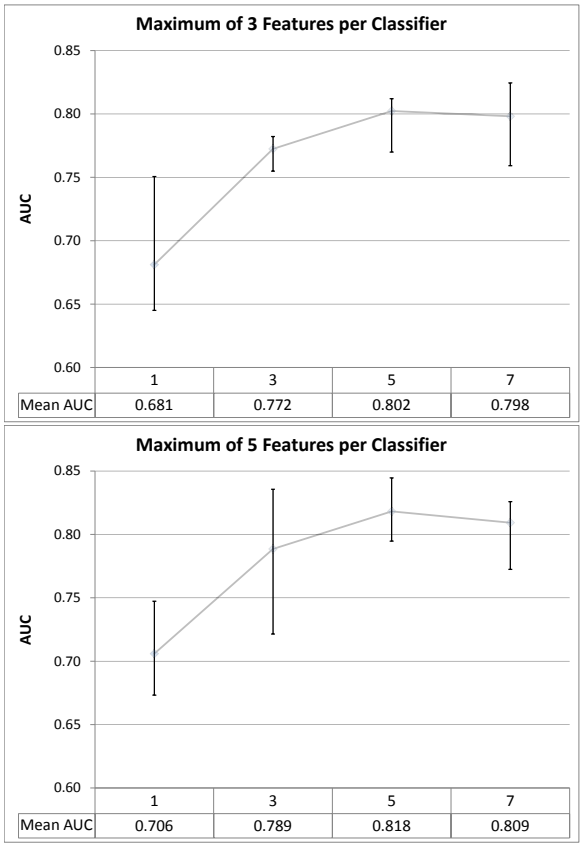

Figure 5: The mean AUC (across 5 trials) vs. the number of classifiers in the ensemble. The end points of the vertical bars denote the maximum and minimum AUCs. Top: Each classifier is limited to 3 features. Bottom: Each classifier is limited to 5 features.

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 during model fusion (model biases complement each other). Figure 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 [13]. 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 [30]. 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 [24], 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 per-

formance 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 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 [13].

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.

References

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

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

[3] L. Breiman, Bagging predictors, Mach. Learn. (1996) 123-140.

[4] L. Buciu, C. Kotropoulos, I. Pitas, Combining support vector machines for accurate face detection, in: Proc. Intl. Conf. on Image Proc. (2001) 1054-1057.

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

[6] O. Chapelle, Training a support vector machine in the primal, Neural Computation 19 (2007) 1155-1178.

[7] O. Chapelle, V. Vapnik, Model selection for support vector machines, Advances in Neural Info. Proc. Systems (1999) 230-236.

[8] O. Chapelle, V. Vapnik, O. Bousquet, S. Mukherjee, Choosing multiple parameters for support vector machines, Mach. Learn. 46 (2002) 131-159.

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

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

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

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

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

[14] S. K. Das, S. Zhou, J. Zhang, F. F. Yin, M. W. Dewhirst, 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 (2007) 1212-1221.

[15] J. O. Deasy, M. Trovo, E. X. Huang, Y. Mu, I. El Naqa, 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 (2008) S119.

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

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

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

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

[20] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, 2002, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (2003) 389-422.

[21] J. Han, M. Kamber, 2006. Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems), 2nd Edition. Morgan Kaufmann.

[22] D. J. Hand, R. J. Till, A simple generalisation of the area under the roc curve for multiple class classification problems, Mach. Learn. 45 (2001) 171-186.

[23] S. Hido, H. Kashima, Roughly balanced bagging for imbalanced data, in: Proc. SIAM Intl. Conf. on Data Mining (2008) 143-152.

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

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

[26] M. Kubat, S. Matwin, Addressing the curse of imbalanced training sets: one-sided selection, in: Proc. Fourteenth Intl. Conf. on Mach. Learn. (1997) 179-186.

[27] J. Li, N. M. Allinson, D. Tao, X. Li, Multitraining support vector machine for image retrieval. IEEE Transactions on Image Processing 15 (2006) 3597-3601.

[28] X. Li, L. Wang, E. Sung, Adaboost with svm-based component classifiers, Engineering Applications of Artificial Intelligence 21 (2008) 785-795.

[29] Z. Liang, Y. Li, Incremental support vector machine learning in the primal and applications, Neurocomputing 72 (2009) 2249-2258.

[30] I. El Naqa, J. D. Bradley, J. O. Deasy, Nonlinear kernel-based approaches for predicting normal tissue toxicities, in: Proc. ICMLA '08 (2008) 539-544.

[31] J. C. Platt, Fast training of support vector machines using sequential minimal optimization, Advances in Kernel Methods - Support Vector Machines (1999) 185-208.

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

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

[34] S. Tong, E. Chang, Support vector machine active learning for image retrieval, in Proc: ACM ICM '01, Multimedia 9 (2001) 107-118.

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

[36] G. Valentini, M. Muselli, F. Ruffino, Cancer recognition with bagged ensembles of support vector machines, Neurocomputing 56 (2004) 461-466.

[37] V. Vapnik, O. Chapelle, Bounds on error expectation for support vector machines, Neural Computation 12 (2000), 2013-2036.

[38] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, V. Vapnik, Feature selection for svms, Advances in Neural Information Processing

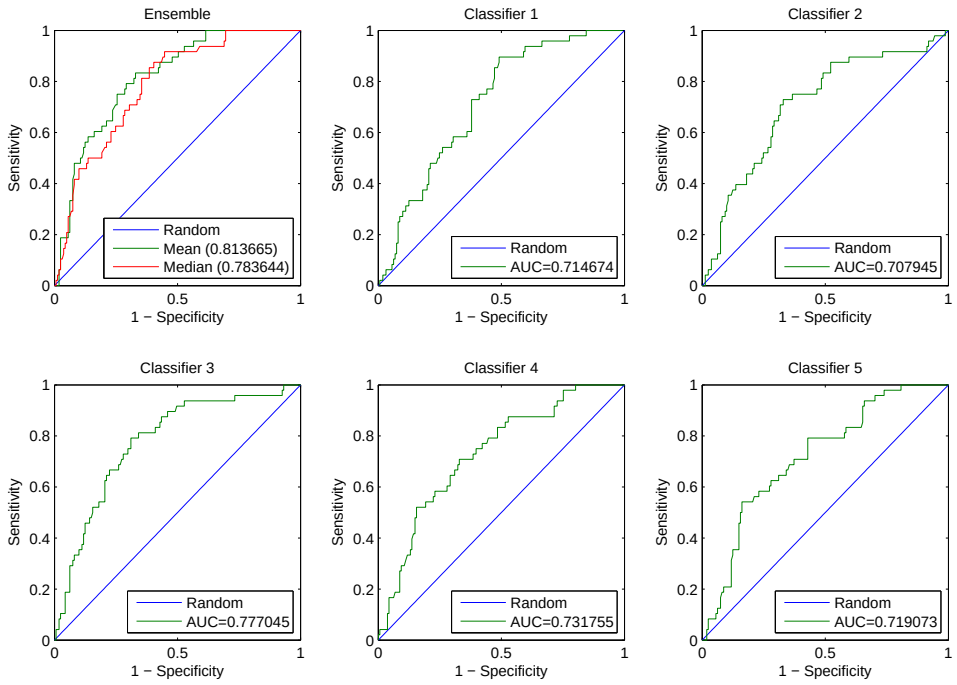

Figure 6: An average performance ensemble classifier with 5 component SVMs each restricted to 5 (the 5/5 model). The relatively weak classifiers complement each other to produce a strong ensemble classifier. The features used by each classifier are shown in Table 2.

Systems 13 (2000) 668-674.

[39] J. Wickramaratna, S. Holden, B. Buxton, Performance degradation in boosting, in: Proc. Second Intl. Workshop on Multiple Classifier Systems 2096 (2001) 11-21.

[40] M. Yamada, S. Kudoh, K. Hirata, T. Nakajima, J. Yoshikawa, Risk factors of pneumonitis following chemoradiotherapy for lung cancer, Eur. J. Cancer 34 (1998) 71-75.

[41] L. Zhang, F. Lin, B. Zhang, Support vector machine learning for image retrieval, in: Proc. ICIP '01 (2001) 721-724.