An Inductive Learning Approach to Prognostic Prediction

W. Nick Street
Departments of Surgery and Computer Sciences
University of Wisconsin
1210 West Dayton Street
Madison, WI 53706
street@cs.wisc.edu

O. L. Mangasarian
Department of Computer Sciences
University of Wisconsin
1210 West Dayton Street
Madison, WI 53706
olvi@cs.wisc.edu

W. H. Wolberg
Department of Surgery
University of Wisconsin
600 Highland Avenue
Madison, WI 53792
wolberg@eagle.surgery.wisc.edu

Abstract

This paper introduces the Recurrence Surface Approximation, an inductive learning method based on linear programming that predicts recurrence times using censored training examples, that is, examples in which the available training output may be only a lower bound on the "right answer." This approach is augmented with a feature selection method that chooses an appropriate feature set within the context of the linear programming generalizer. Computational results in the field of breast cancer prognosis are shown. A straightforward translation of the prediction method to an artificial neural network model is also proposed.

1 INTRODUCTION

Machine learning methods have been successfully applied to the analysis of many different complex problems in recent years, including many biomedical applications. One field which can benefit from this type of approach is the analysis of *survival* or *lifetime data* (Lee, 1992; Miller Jr., 1981), in which the objective can be broadly defined as predicting the time of a particular event. In this work we are concerned specifically with prognosis, that is, predicting the course of a disease. This paper introduces a new prognosis prediction method, based on linear programming (Dantzig, 1963), which can be implemented using a backpropagation neural networks (Rumelhart et al., 1986) based on the same objective. These methods are applied to breast cancer prognosis, predicting how long after surgery we can expect the disease to recur. We also describe a feature selection method which has been incorporated into our linear model, and show that the evaluation of the effectiveness of the various input features is medically important.

In applying inductive machine learning to prognosis, we first note that this problem does not fit into either of the classic paradigms of classification or function approximation. While a patient can be classified "recur" if the disease is observed, there is no real cutoff point at which the patient can be considered a non-recurrent case. The data are therefore *censored* in that we know a time to recur (TTR) for only a subset of patients; for the others, we know only the time of their last checkup, or disease free survival time (DFS). In particular, recurrence or survival data is *right censored*, i.e., the right endpoint (recurrence time) is sometimes unknown, since some patients will inevitably move away, change doctors, or die of unrelated causes.

Problems involving censored data are common to several fields. In engineering, one might be interested in the survival characteristics of electronic components, while sociologists might consider what factors lead to long-lasting marriages. However, the application of machine learning methods to these problems has been rare. Schenone *et al.* (Schenone et al., 1993) used a self-organizing neural network to find classes of cases with similar expected recurrence times. However, they did not directly address the problem of using censored data, that is, cases which have not been followed to recurrence / death. Burke (Burke, 1994) used artificial neural networks (ANNs) to approach prognosis as a separation problem, as was done in previous work at Wisconsin (Wolberg et al., 1992; Wolberg et al., 1994). This is done by choosing one or more endpoints and separating sets such as "patients who recurred in less than two years." The work of Ravdin and colleagues (De Laurentiis and Ravdin, 1994; Ravdin and Clark, 1992; Ravdin et al., 1994) used ANNs to generate survival curves, which plot the probability of disease-free survival against time. This work uses the the trained network's output as an approximation of recurrence probability. While their cumulative results closely fit the population recurrence characteristics of the test cases, we believe that better individual predictions may be obtained by directly predicting survival time.

2 RECURRENCE SURFACE APPROXIMATION

2.1 LINEAR PROGRAMMING FORMULATION

We approach the prediction of time to recur (TTR) as a function estimation problem, a mapping of an n-dimensional input of cytological and other features to a one-dimensional time output. Complicating the problem is the fact that TTR is known for only a subset of patients; for the others, we know only the time of their last check-up, or disease-free survival time (DFS). The most obvious approach to predicting TTR is to use only the uncensored cases – those for which a TTR is known – and use the input features to fit them with least squares (or other least-error) procedure. However, by exploiting the straightforward manner in which inequalities are handled in linear programming, we are able to include all available cases to build a more accurate, robust predictive model.

Our solution to this estimation problem is termed the recurrence surface approximation (RSA) technique (Mangasarian et al., 1994; Street, 1994; Wolberg et al., 1995). RSA uses linear programming to determine a linear combination of the input features that accurately predicts TTR. The intuitive motivation for the RSA approach is that:

- Recurrences actually take place at some point in time prior to their detection. However, the difference between the time a recurrence is detectable (actual TTR) and the time it is actually detected (observed TTR) is assumed to be small.
- Observed disease-free survival time (DFS) is a *lower bound* on the recurrence time of that patient.

These assumptions can be formulated into the following linear program for a given training set:

$$\begin{array}{lrcccl} \underset{w,\gamma,v,y,z}{\text{minimize}} & \multicolumn{5}{l}{\frac{1}{m}e^T y + \frac{1}{k}e^T z + \frac{\delta}{m}e^T v} \\ & -v & \leq & Mw + \gamma e - t & \leq & y \\ \text{subject to} & & & -Nw - \gamma e + r & \leq & z \\ & & & v, y, z & \geq & 0 \end{array} \qquad (1)$$

The purpose of this linear program is to learn the weight vector w and the constant term γ. These parameters determine a recurrence surface $s = xw + \gamma$, where x is the n-dimensional vector of measured features and s is the surface (in this case, a plane defined on the feature space) that predicts recurrence times. Here M is an $m \times n$ matrix of the m recurrent points, with recurrence times given by the m-dimensional vector t. Similarly, the k non-recurrent points are collected in the $k \times n$ matrix N, and their last known disease-free survival times are in the k-dimensional vector r. The vectors y and z represent the errors for recurrent and non-recurrent points, respectively; overestimating the TTR of recurrences is considered an error, while predicting a TTR smaller than an observed DFS is also an error. The objective averages the errors over their respective classes. (Note: e is a vector of 1's of appropriate dimension.) The linear program is implemented using the MINOS numerical optimization software (Murtagh and Saunders, 1983).

Because recurrences take place at some unknown time prior to their detection, we do not consider underestimated recurrent points to be as serious of an error as overestimated ones. To reflect this, the v term in the objective is weighted by an appropriately small positive parameter δ, forcing underestimated recurrent points closer to the surface. Based on a perturbation theorem (Mangasarian and Meyer, 1979), for a sufficiently small positive δ, that is $0 < \delta \leq \bar{\delta}$ for some $\bar{\delta}$, the objective minimizes the weighted term conditionally, i.e., of those possible variable values which minimize the first two terms of the objective, those values which minimize the third term are chosen. In this work, the "sufficiently small" value of δ was chosen empirically, by lowering it until further reductions had no effect on the training objective.

One possible variation of the RSA linear formulation is to vary the relative importance of the recurrent and non-recurrent points. The above approach considers the two *sets* of points to be equally important by averaging them separately. One straightforward modification which has been tested is to weight each example equally, resulting in the following *pooled error* objective:

$$\frac{1}{m+k}e^T y + \frac{1}{m+k}e^T z + \frac{\delta}{m+k}e^T v \qquad (2)$$

In fact, varying the relative weight applied to the recurrent and non-recurrent points is an easy way to change the predictive characteristics of the resulting generalizer, depending on the application. For instance, if most cases are uncensored, the censored cases may add only slightly more information and could be appropriately weighted with a small multiplier.

2.2 FEATURE SELECTION

Overfitting avoidance can sometimes be achieved by training on only a subset of the given features. This is particularly true when the number of features is large relative to the number of training cases. This section describes a simple, automatic method which incorporates feature selection into the linear programming learning algorithm, thereby explicitly considering the biases inherent in the learning procedure in the search for a useful feature subset.

Essential to the feature-selection algorithm is the fact that the features have all been normalized to lie in approximately the same range. All ordered features – whether real valued or integer valued – are "z-transformed" to have mean zero and standard deviation one. Since we are constructing linear models, wherein the predictions are constructed via a linear combination of the input features, this allows the assumption that the relative importance of an individual feature in the predictive model is approximated by the magnitude of the feature's corresponding linear coefficient. Note that this scaling has two useful side effects:

- Outlier feature values (that is, individual values that are extremely large or small relative to the mean value of the feature) are easily identified, since the value reflects the number of standard deviations the original value is from the mean. Outliers may have some domain-specific semantics and may require special handling.
- Missing feature values can be safely set to zero. Setting missing values to the mean is a common practice which is particularly appropriate here. Since we are building linear models, a missing value has no effect on the prediction for that case.

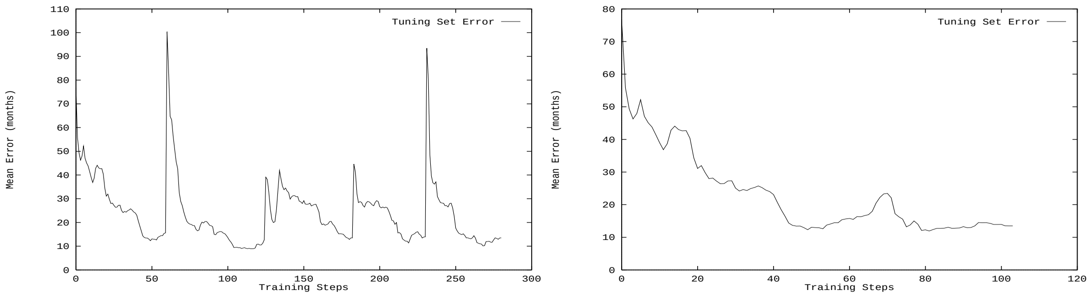
Our approach is similar to John *et al.* (John et al., 1994) in that we employ a "wrapper model", incorporating the selection of features and evaluation of the resulting subset into the learning algorithm itself. The feature-selection procedure is a variation of the heuristic sequential backward elimination method (Kittler, 1986; Marill and Green, 1963), a top-down, greedy search through the space of feature sets. The procedure begins by setting aside a tuning or validation set (Lang et al., 1990), that is, a surrogate testing set, in our case a randomly selected 20% of the training cases. The regular RSA procedure is then applied to the training set, finding the global minimum of the training objective using all the available features. We then examine the resulting weight vector w, and force the following weights to zero: all those which are non-basic (zero) variables in the LP solution, and the smallest in magnitude among the (usually non-zero) weights represented by basic variables. The RSA LP is then reformulated and a new solution found in this reduced dimension. This elimination of variables is repeated until only one variable remains. For the final predictive model, we choose the feature set which demonstrated the best performance on the tuning examples. Since setting aside a tuning set can cause significant degradation when using only a small number of samples, we add the tuning set back into the training set, and perform one final optimization step.

Figure 1 shows the observed error on the tuning set for a typical run. In Figure 1(a), each reduction of the problem dimension results in a jump in the error, which is then reduced through the optimization process. However, by carefully utilizing the "hot start" capability of the MINOS linear programming package, these error spikes can be eliminated. Eliminating a variable can be accomplished by setting its upper and lower bounds both to zero. Since only the bounds have changed, the new linear program can be solved starting at the solution to the previous LP. Since the solution in the new space typically lies very close to the previous solution, these re-optimizations are very fast. As shown in Figure 1(b), the number of training iterations is substantially reduced using hot starts.

2.3 APPLICATION TO WISCONSIN PROGNOSTIC DATA

In previous work (Street et al., 1993; Wolberg et al., pted; Wolberg et al., 1993; Wolberg et al., 1994) the authors developed an image-processing software package for breast cancer diagnosis, known as **Xcyt**, which analyzes digital images of cells taken from breast lumps. This program computes 30 different features of the cellular nuclei in the image, including measurements of size, shape and texture. Since its incorporation into clinical practice in 1993, a diagnostic classification system built with a subset of these features has correctly classified all of the 136 consecutive new cases tested in that time as benign or malignant.

In the present application, the input consists of all 30 nuclear features computed by **Xcyt** together with two traditional prognostic predictors: tumor size and number of involved lymph nodes. The prognosis data set consists of those malignant patients for which follow-up data was available, after eliminating those cases with distant metastasis (cancer has already spread;



(a) Tuning set error: Cold start

(b) Tuning set error: Hot start

Figure 1: The tuning-set error during training with feature-set selection. In part (a), each spike corresponds to a dropped feature, while the subsequent drop in error corresponds to re-optimization in the reduced feature space. In part (b), the error spikes associated with dropping a feature have been eliminated and the number of training steps has been reduced significantly.

prognosis is poor) and carcinoma *in situ* (cancer has not yet invaded breast tissue; prognosis is good). This left 187 cases, 44 of which have recurred.

The RSA procedure was tested with leave-one-out testing (Lachenbruch and Mickey, 1968) to evaluate its accuracy in predicting future outcomes. Table 1 shows the mean generalization errors of the RSA formulation compared with the following prediction methods:

- **Pooled RSA:** All of the feature examples were weighted equally in the objective. See Equation 2.
- **Least 1-norm error on recurs:** An obvious method for predicting recurrence is to construct the predictive surface by a least-error fit on those cases for which the outcome is known, in our case, those with an observed recurrence time. We chose the 1-norm error, minimizing the average error on the recurrent cases but testing on all the examples. For compatibility, this test was run using the greedy feature selection method described earlier.
- **k-nearest neighbors:** The nearest neighbor procedure (Aha et al., 1991; Hart, 1967) is another effective and intuitive method for generalization. In this application, the recurrence time for the k recurrent cases closest to the given test point in Euclidean space were averaged to give a prediction. We tested values of k between one and ten, with both scaled and unscaled features. In order to simulate feature selection, the k-nearest neighbor algorithm was also tested using only the six features found most relevant by RSA (see previous section). The best results, which are reported in Table 1, were obtained using all of the unscaled features, and k equal to seven.

Comparative results on all points, recurrent cases only and non-recurrent cases only are shown in Table 1 for the various prediction methods. We emphasize that these are estimates of the method's real-world performance, and measure only those cases known to be in error: overestimated recurrences (prediction was late) and underestimated non-recurrent cases (prediction was early). Both RSA approaches significantly lower the mean prediction error compared to a simple fit. Using a pairwise t-test, predictions of the original RSA formulation are better than those of the nearest neighbor routine at the 95% confidence level, and better than those of the 1-norm error fit at the 99% confidence level. We note that the original RSA formulation performs comparably on both recurrent and non-recurrent cases, while the pooled error method greatly favors the majority non-recurrent class, thereby lowering the mean error. However, since our goal is to predict *all* cases as accurately as possible, the original RSA formulation is considered superior. As expected, the recurrent results for the least 1-norm estimation are good, while that method does worst on the non-recurs and has a high overall mean error. This is attributable to the fact that most of the observed recurrences were at relatively short times (the mean recurrence time was 24 months), therefore a regression method which uses only the recurrent cases skews the predictions downward, matching the bias of this particular data set. The nearest-neighbor predictor similarly benefitted from this bias, predicting the re-

Table 1: Average error (months) of various prognostic formulations on Wisconsin prognostic data using leave-one-out testing.

Predictive Formulation	Average Error (months)		
	All Points	Non-recur	Recur
RSA (Eq. 1)	18.3	19.9	13.0
Pooled RSA (Eq. 2)	13.9	6.1	39.3
Least 1-norm Error	23.2	27.2	10.4
k-Nearest Neighbor	21.5	25.3	9.1

current cases well.

The recurrence predictions made by RSA can also be interpreted using survival curves. Again using the leave-one-out testing data, the cases were divided into three groups based upon their predicted time of recurrence: predictions of less than two years (54 cases), predictions from two to five years (98 cases), and predictions greater than five years (35 cases). The actual recurrence probabilities of these three groups were then constructed with the Kaplan-Meier (Kaplan and Meier, 1958) approximation, as shown in Figure 2.

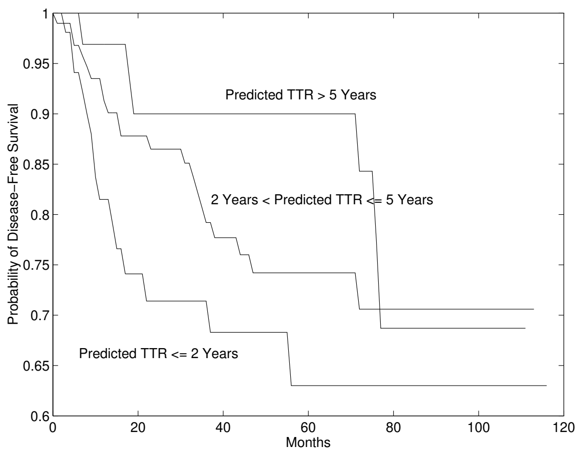


Figure 2: Kaplan-Meier survival probabilities based on RSA predicted TTR.

This plot clearly shows that the RSA predictions for these cases was consistent with the actual prognosis. Those examples for which the prediction was less than two years have the worst prognosis for the duration of the study, and a recurrence rate of nearly 30% at two years. Those with predicted TTR's of greater than five years have an extremely good prognosis, with only 10% recurrence up to five years, but with more recurrences being observed (as expected) after that time.

Implicit in all of the above results is the use of the feature selection method described in Section 2.2. Table 2 summarizes the computational effect of feature selection, performing the RSA procedure on the Wisconsin prognostic data. As in Table 1, the cumulative error is divided into errors in predicting recurrent and non-recurrent cases. Three different styles of feature selection were tested: no selection, tuning-set selection as described above, and a variation which eliminates features down to a certain predetermined number. In this case, a cross-validation determined that the average number of features required was four. All results reflect leave-one-out testing.

In all cases, eliminating redundant or irrelevant features improved the results significantly. Using cross-validation to determine the number of features in each model may also improve performance, by reducing the variation allowed by tuning with a small subset. Further, finding the most important features is itself an important goal, particularly in medical domains such as prognosis, since that particular piece of information may have been obtained through an expensive test or a potentially dangerous procedure. We will re-examine this point in Section 3.

2.4 APPLICATION TO OTHER DATASETS

2.4.1 SEER Data

The RSA prognostic prediction method was tested using the breast database from the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (Carter et al., 1989), which contains follow-up data for over 44,000 breast cancer patients. These cases contain five integer-valued input features, as follows:

1. Histological Grade: 1, 2, or 3.
2. Tumor Size: 1, 2, ..., or 11.
3. Tumor Extension: 1, 2, ..., or 5.
4. Number of Axillary Lymph Nodes Positive: 1, 2, ..., or 10.

Table 2: Average error (months) of RSA formulation with various feature selection methods: Wisconsin prognostic data.

	Mean Error	Non-recur Error	Recur Error
Original RSA	21.8	20.1	27.5
RSA: Tuned Selection	18.3	19.9	13.0
RSA: Select 4	17.7	19.0	13.6

5. Number of Axillary Nodes Examined: 1, 2, ... or n.

Although the values are coded, they typically follow the "bigger means worse" ordering which was coded into the Wisconsin features. There is a high incidence of missing feature values in this data; for instance, histological grade is recorded for only about 17% of the cases.

Table 3 reports similar results for different subsets of the SEER database. As this database of 44,135 cases produced models that were too large for the memory of our workstations, a subset of 1,000 cases was selected by choosing every 44th record. After again eliminating cases with distant metastasis and *in situ* cancer, 890 remained for testing. Because of the small number of features, the feature selection had little effect on the results, so only tests using it will be reported. A second test was performed with just the 7,438 cases which included a value for histological grade. Note that the SEER data records survival time as the endpoint rather than recurrence time.

The errors on this larger data set confirm the effectiveness of the RSA method. We have also shown that a precise, detailed description of nuclear grade is enough information to predict prognosis as well as the more traditional features recorded in the SEER study. However, histological grade seems to have little effect on the predictive power of the SEER model. This we attribute to the small amount of information available in an objectively graded, three-valued feature.

2.4.2 Gutenberg Data

We also evaluated the RSA procedure on a third data set from the group at Gutenberg, Germany (Mitze, 1992). These cases contain 13 integer and coded features: age at diagnosis, menopausal status, histologic grading, number of lymph nodes removed, number of lymph nodes involved, tumor size, estrogen receptor status, progesterone receptor status, immunohistologic overexpression of erbB-2, urokinase-like plasminogen activator, proliferation rate via ki67 antigen, plasminogen activatou inhibitor (PAI-1), and Cathepsin D. Computational results with these cases are summarized in Table 4.

First, these positive results further validate the use of the RSA method for prognosis prediction. However, adding the feature selection procedure did not help in this case, as the risk of eliminating useful features outweighed the benefit of overtraining avoidance.

The average errors are significantly lower here than with the other data sets. However, we attribute this less to the utility of the Gutenberg features than to the relative homogeneity of the cases. No non-recurrent cases with follow-up time less than 24 months were included. Further, 66% of the recurrence times and 94% of the follow-up times are between two and six years. Hence, this prediction task was significantly easier than the other two explored.

The feature usage on the "select five" experiment also supports the above conclusion. The most prevalent feature in the 141 testing cases was histological grade, which was selected 97% of the time. Other relevant features were involved lymph nodes (chosen in 87% of the tests) and estrogen receptor status (72%). Since the other two data sets also involved some indication of nuclear grading and lymph node status, we conclude that the new features such as the hormone-related measurements were not the source of the improved results.

3 CLINICAL APPLICATION

Both medical and personal decisions hinge on the projected future course of the breast cancer. Decisions whether or not chemotherapy is needed and the intensity of such therapy are based on the anticipated course of the cancer. The mental state of the patient, and personal and career plans are greatly affected by the anticipated course of the disease. Hence, improved prognostic prediction is an important goal for cancer treatment.

The best predictive model using the RSA formulation (Equation 1) has been added to the **Xcyt** program. We determined (using cross-validation) that five input

Table 3: Average error (in months) of RSA: SEER data.

	Mean Error	Non-recur Error	Recur Error
RSA: Random Cases	17.2	17.0	18.0
RSA: Histo. Grade Cases	17.6	18.1	16.4

Table 4: Average error (in months) of RSA formulations: Gutenberg data.

	Mean Error	Non-recur Error	Recur Error
Original RSA	6.36	5.56	9.08
Pooled RSA	5.70	2.79	15.61
Original: Tuned Selection	8.01	7.88	8.46
Pooled: Tuned Selection	5.68	2.43	16.73
Original: Select 5	7.43	7.28	7.91
Pooled: Select 5	5.91	2.56	17.28

features were needed for this prediction task. [1] We then used the above feature-selection scheme to pare down the original set of 32 features to five: mean value of radius, perimeter, and fractal dimension, and extreme (largest)value of perimeter and area. Using this model, we now predict a time of recurrence for patients who have been diagnosed with a malignant tumor, to aid the choice of treatment. Further, **Xcyt** provides estimates of the patient's probability of disease-free survival in the form of a Kaplan-Meier curve (Figure 3). The disease-free-survival curve for the individual patient is based on those training cases that had a similar predicted time of recurrence as determined by the RSA (1).

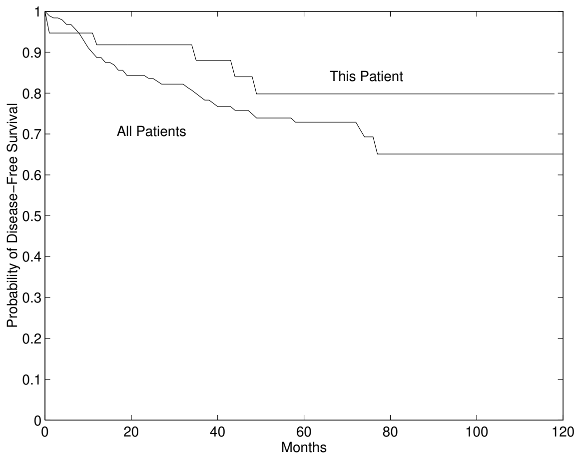


Figure 3: RSA-based Kaplan-Meier estimated probability of disease-free survival, up to ten years following surgery, for all patients in the study and for one particular case.

The feature-selection procedure also provides insight into which of the input features is most important for prognosis prediction. Consider the "select four" test run with the original RSA formulation, in which leave-one-out testing results in 187 separate predictive models. The mean radius feature was used in 155 (83%) of those models, followed by worst radius (145), worst area (132) and mean perimeter (129). In fact, the nine size-related variables (of the original set of 32 variables) accounted for 91% of the total feature usage. Clearly, nuclear size is significantly more important for prognostic estimation than any of our representations of shape. It is also interesting to note that the traditional tumor size and lymph node status features were used in none of the 187 tests.

In current traditional medical practice, the strongest available prognostic feature is the extent to which cancer is present in the lymph nodes, which is determined by microscopic examination of lymph nodes that must be surgically removed from the patient's armpit. This procedure leaves the patient more susceptible to infection and the arm frequently develops lymphedema, a potentially severe swelling of the arm (Aitken et al., 1989; Kissin et al., 1986). Additionally, prognostic determinations based on lymph node involvement are inaccurate: 10% of patients in the most favorable category will die of breast cancer and 40% of those in the most unfavorable category will survive. Therefore, one of the most important findings of this research is that the best predictive models are found using just nuclear features, and are not improved by the traditional medical prognostic factors of tumor size and lymph node

[1] This result uses an updated version of the Wisconsin database, resulting in slightly different feature usage.

status. If further studies confirm these findings, the routine and potentially hazardous and debilitating removal of lymph nodes from the armpit of breast cancer patients for prognostic purposes can be avoided.

4 CONCLUSIONS AND FUTURE WORK

We have presented the Recurrence Surface Approximation method, an effective approach to the prediction of lifetime or recurrence. Through our application to breast cancer, we can provide comparatively accurate, patient-specific predictions of when the cancer is likely to recur. By using a heuristic feature selection method, we are able to identify important prognostic factors, an important clinical task. Because it uses censored data to build a predictive survival model, the RSA method is applicable to many different fields. The potential for applying these same approaches to other medical decision making, prediction and machine learning problems appears to be extremely promising and is worthy of further investigation and testing. In particular, we are preparing an extensive set of tests comparing our approach to techniques from the statistical literature, such as Cox proportional hazards regression (Cox, 1972).

So far only linear models have been employed by the RSA method, which limits its representational power. We therefore plan to modify backpropagation artificial neural networks (ANN's) (Rumelhart et al., 1986) to predict recurrence time in the same manner as the RSA procedure. The output node of the network will use the identity function as its transfer function, so that it can learn any floating point number. The error function at the output node will be changed to a squared-error version of the RSA error:

$$e = \begin{cases} (target - predicted)_+^2 & \text{censored case} \\ (predicted - target)_+^2 & \text{non-censored case} \\ \quad +\delta(target - predicted)_+^2 & \end{cases}$$

where e is the error on a particular example, $x_+ := max\{x, 0\}$ and δ is again a small positive constant.

Another important research direction is the use of the RSA method to evaluate the relative effectiveness of different treatment strategies. By building separate prognostic prediction surfaces using patients who received different treatments, we can determine which regions of the feature space correspond to an improved prognosis for each particular treatment. The portion of the feature space where one treatment is clearly favored would represent a patient profile, a set of characteristics which would indicate use of that treatment. As with all of our medical applications, the goal is to provide doctors and patients with predictive models that increase the accuracy with which clinicians can plot the future course of diseases, resulting in treatment decisions in which both doctor and patient have the most reliable information.

Acknowledgements

The authors would like to thank Don Henson and Margarete Mitze for use of the SEER and Gutenberg databases, respectively. This research was supported by Air Force Office of Scientific Research Grant F–49620–94–1–0036 and National Science Foundation Grant CCR–9322479.

References

Aha, D. W., Kibler, D., and Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, 6:37–66.

Aitken, R. J., Gaze, M. N., Rodger, A., Chetty, U., and Forrest, A. P. M. (1989). Arm morbidity within a trial of mastectomy and either nodal sample with selective radiotherapy or axillary clearance. *British Journal of Surgery*, 76:568–571.

Burke, H. B. (1994). Artificial neural networks for cancer research: Outcome prediction. *Seminars in Surgical Oncology*, 10:73–79.

Carter, C. L., Allen, C., and Henson, D. E. (1989). Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. *Cancer*, 63:181–187.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society*, B 34:187–202.

Dantzig, G. B. (1963). *Linear Programming and Extensions*. Princeton University Press, Princeton NJ.

De Laurentiis, M. and Ravdin, P. M. (1994). A technique for using neural network analysis to perform survival analysis of censored data. *Cancer Letters*, 77:127–138.

Hart, P. (1967). The condensed nearest neighbor rule. *Transactions on Information Theory*, IT-14:515–516.

John, G. H., Kohavi, R., and Pfleger, K. (1994). Irrelevant features and the subset selection problem. In *Proceedings of the 11th International Conference on Machine Learning*, San Mateo, CA. Morgan Kaufmann.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481.

Kissin, M. W., Querci della Rovere, G., Easton, D., and Westbury, G. (1986). Risk of lymphoedema following the treatment of breast cancer. *British Journal of Surgery*, 73:580–584.

Kittler, J. (1986). Feature selection and extraction. In Fu, Y. ., editor, *Handbook of Pattern Recognition and Image Processing*. Academic Press, New York.

Lachenbruch, P. and Mickey, P. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, 10:1–11.

Lang, K., Waibel, A., and Hinton, G. (1990). A time-delay neural network architecture for isolated word recognition. *Neural Networks*, 3:23–43.

Lee, E. T. (1992). *Statistical Methods for Survival Data Analysis*. John Wiley and Sons, New York.

Mangasarian, O. L. and Meyer, R. R. (1979). Nonlinear perturbation of linear programs. *SIAM Journal on Control and Optimization*, 17:745–752.

Mangasarian, O. L., Street, W. N., and Wolberg, W. H. (1994). Breast cancer diagnosis and prognosis via linear programming. Technical Report 94–10, University of Wisconsin. *Operations Research*, accepted. Available from ftp://ftp.cs.wisc.edu/math-prog/tech-reports/.

Marill, T. and Green, D. M. (1963). On the effectiveness of receptors in recognition systems. *IEEE Transactions on Information Theory*, 9:11–17.

Miller Jr., R. G. (1981). *Survival Analysis*. John Wiley and Sons, New York.

Mitze, M. (1992). Personal communication.

Murtagh, B. and Saunders, M. (1983). MINOS 5.0 user's guide. Technical Report SOL 83.20, Stanford University. MINOS 5.4 Release Notes, December 1992.

Ravdin, P. M. and Clark, G. M. (1992). A practical application of neural network analysis for predicting outcome of individual breast cancer patients. *Breast Cancer Research and Treatment*, 22:285–293.

Ravdin, P. M., Tandon, A. K., Allred, D. C., Clark, G. M., Fuqua, S. A. W., Hilsenbeck, S. H., Chamness, G. C., and Osborne, C. K. (1994). Cathepsin D by western blotting and immunohistochemistry: Failure to confirm correlations with prognosis in node-negative breast cancer. *Journal of Clinical Oncology*, 12:467–474.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by error propagation. In Rumelhart, D. E. and McClelland, J. L., editors, *Parallel Distributed Processing*, volume 1, chapter 8. MIT Press, Cambridge, MA.

Schenone, A., Andreucci, L., Sanguinetti, V., and Morasso, P. (1993). Neural networks for prognosis in breast cancer. *Physica Medica*, IX(Supplement 1):175–178.

Street, W. N. (1994). *Cancer Diagnosis and Prognosis via Linear-Programming-Based Machine Learning*. PhD thesis, University of Wisconsin-Madison. Available as University of Wisconsin Mathematical Programming TR 94–14 from ftp://ftp.cs.wisc.edu/math-prog/tech-reports/.

Street, W. N., Wolberg, W. H., and Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis. In *IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology*, volume 1905, pages 861–870, San Jose, California.

Wolberg, W. H., Bennett, K. P., and Mangasarian, O. L. (1992). Breast cancer diagnosis and prognostic determination from cell analysis. Manuscript, Departments of Surgery and Human Oncology and Computer Sciences, University of Wisconsin.

Wolberg, W. H., Street, W. N., Heisey, D. H., and Mangasarian, O. L. (1995). Computer-derived nuclear grade and breast cancer prognosis. *Analytical and Quantitative Cytology and Histology*. accepted.

Wolberg, W. H., Street, W. N., Heisey, D. H., and Mangasarian, O. L. (accepted). Computer-derived nuclear features distinguish malignant from benign breast cytology. *Human Pathology*.

Wolberg, W. H., Street, W. N., and Mangasarian, O. L. (1993). Breast cytology diagnosis via digital image analysis. *Analytical and Quantitative Cytology and Histology*, 15(6):396–404.

Wolberg, W. H., Street, W. N., and Mangasarian, O. L. (1994). Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. *Cancer Letters*, 77:163–171.