Survival-Time Classification of Breast Cancer Patients

Y.-J. Lee [*], O. L. Mangasarian [†]& W. H. Wolberg [‡]

Keywords breast cancer, support vector machines, classification

Dedication (Mangasarian) To Lucien Polak, friend, colleague and a major contributor to mathematical programming, on the occasion of his 72nd birthday.

Abstract

The identification of breast cancer patients for whom chemotherapy could prolong survival time is treated here as a data mining problem. This identification is achieved by clustering 253 breast cancer patients into three prognostic groups: Good, Poor and Intermediate. Each of the three groups has a significantly distinct Kaplan-Meier survival curve. Of particular significance is the Intermediate group, because patients with chemotherapy in this group do better than those without chemotherapy in the same group. This is the reverse case to that of the overall population of 253 patients for which patients undergoing chemotherapy have worse survival than those who do not. We also prescribe a procedure that utilizes three nonlinear smooth support vector machines (SSVMs) for classifying breast cancer patients into the three above prognostic groups. These results suggest that the patients in the Good group should not receive chemotherapy while those in the Intermediate group should receive chemotherapy based on our survival curve analysis. To our knowledge this is the first instance of a classifiable group of breast cancer patients for which chemotherapy can possibly enhance survival.

[*]Department of Computer Science & Information Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan 106, *yuh-jye@mail.ntust.edu.tw*.

[†]Computer Sciences Department, University of Wisconsin, Madison, WI 53706, *olvi@cs.wisc.edu*.

[‡]General Surgery, Clinical Sciences Center, University of Wisconsin, Madison, WI 53792, *wolberg@surgery.wisc.edu*.

1 Introduction

In this work we introduce an important application of data mining to breast cancer prognosis. The principal objective of this work is to try to identify breast cancer patients for whom chemotherapy prolongs survival time. By chemotherapy we mean adjuvant chemotherapy, that is chemotherapy administered shortly following the initial surgery when the patient has no evidence of distant metastatic disease. This should be distinguished from chemotherapy that is given when the patient has evidence of distant metastatic disease at some later time, that is when cancer has spread to other organs of the body. Because we cannot carry out comparative tests on human subjects, similar breast cancer patients must be treated similarly. Here the similarity is based on physicians' knowledge. In our approach instead, we utilize data mining techniques such as support vector machine classification, feature selection and clustering to identify a group of patients who could benefit from chemotherapy. This identification is achieved by clustering 253 breast cancer patients listed in the publicly available [19] WPBCC dataset into three prognostic groups: a Good group consisting of 69 patients all without chemotherapy and which collectively have the best survival curve; a Poor group consisting of 73 patients all with chemotherapy and which collectively have the worst survival curve; and an Intermediate group consisting of 44 patients without chemotherapy and 67 patients with chemotherapy. Each of the three groups has a significantly distinct Kaplan-Meier survival curve [6, 7]. Of particular significance is the Intermediate group, because patients with chemotherapy do better than those without chemotherapy in this group. This is the reverse case to that of the overall population of 253 patients for which patients undergoing chemotherapy have worse survival than those who do not.

We use a support vector machine classification procedure to classify each patient into one of these three groups. Because of the complexity of this *multicategory* classification problem, a simple application of even a nonlinear support vector machine does not yield satisfactory test set correctness. Instead we perform a preliminary classification in a 6-feature space consisting of 5 cytological features (mean of area, standard error of area, worst area, worst texture and worst of perimeter) [17, 16] and one pathology feature (tumor size) tumor size. We then compute 2 additional dependent features based on this preliminary separation and use the combined 8 features (6 original ones and 2 dependent ones) to achieve our final separation. We are able to achieve an 82.7% test set correctness using this classification procedure.

The final result is the Prognostic Procedure 4.1 that assigns a new patient to one of three groups without making use of the lymph node status. Here by lymph node status we mean the number of metastasized lymph nodes. Lymph nodes are removed during surgery in conjunction with the removal of the malignant tumor from the breast. This potentially risky procedure which can cause arm swelling and increased susceptibility to infection can be eliminated by using the classification procedures we propose here.

The paper is organized as follows. In Section 2 we introduce the support vector machine that will be used in the classification process of Section 4 of the paper. An efficient computational algorithm, the smooth support vector machine (SSVM) [9] is also described in this section of the paper that implements the generation of a support vector machine classifier. Section 3 describes our clustering procedure which generates the three survival groups while Section 4 implements our approach and uses the SSVM algorithm to generate a nonlinear SVM procedure to classify the patients into three survival groups. Section 5 concludes the paper.

A word about background material and notation. Kaplan Meier survival curves [6, 7], used extensively in quantifying survival rates, give the percentage of surviving patients as a function of time. Two survival curves are considered to be distinct by the log-rank statistic if the p-value is less than 0.05 [7]. Turning to our notation, all vectors will be column vectors unless transposed to a row vector by a prime superscript $'$. For a vector x in the n-dimensional real space R^n, the plus function x_+ is defined as $(x_+)_i = x_i$ if $x_i > 0$ else $(x_+)_i = 0$ if $x_i \leq 0$ for $i = 1, \ldots, n$. The scalar (inner) product of two vectors x and y in the n-dimensional real space R^n will be denoted by $x'y$ and the 2-norm of x will be denoted by $\|x\|_2$. For a matrix $A \in R^{m\times n}$, A_i is the ith row of A which is a row vector in R^n. A column vector of ones of arbitrary dimension will be denoted by e. For $A \in R^{m\times n}$ and $B \in R^{n\times k}$, the kernel $K(A, B)$ maps $R^{m\times n} \times R^{n\times k}$ into $R^{m\times k}$. In particular, if x and y are column vectors in R^n then, $K(x', y)$ is a real number, $K(x', A')$ is a row vector in R^m and $K(A, A')$ is an $m \times m$ matrix. The base of the natural logarithm will be denoted by ε. We will make use of the following Gaussian kernel [18, 5, 11] that is frequently used in the SVM literature:

$$K(A, A')_{ij} = \varepsilon^{-\mu\|A_i - A_j\|_2^2},\ i = 1 \ldots, m,\ j = 1 \ldots, m, \tag{1}$$

where $A \in R^{m\times n}$ and μ is a positive constant usually determined by a tuning set.

2 The Smooth Support Vector Machine (SSVM)

In this section of the paper we give a brief description of SSVM [9] that is used for our classification procedure. We consider the problem of classifying m points in the n-dimensional real space R^n, represented by the $m \times n$ matrix A, according to membership of each point A_i in the classes +1 or -1 as specified by a given $m \times m$ diagonal matrix D with ones or minus ones along its diagonal. For this problem the standard support vector machine with a linear kernel AA' [18, 5] is given by the following for some $\nu > 0$:

$$\begin{array}{rl} \min\limits_{(w,\gamma,y)\in R^{n+1+m}} & \nu e'y + \frac{1}{2}w'w \\ \text{s.t.} & D(Aw - e\gamma) + y \geq e \\ & y \geq 0. \end{array} \tag{2}$$

Here w is the normal to the bounding planes:

$$x'w - \gamma = \pm 1 \tag{3}$$

and γ determines their location relative to the origin. The first plane above bounds the class +1 points and the second plane bounds the class -1 points when the two classes are strictly linearly separable, that is when the slack variable $y = 0$. The linear separating surface is the plane:

$$x'w = \gamma, \tag{4}$$

midway between the bounding planes (3). See Figure 1. If the classes are linearly inseparable then the two planes bound the two classes with a "soft margin" determined by a nonnegative slack variable y, that is:

$$\begin{array}{rcrcrcl} x'w & - & \gamma & + & y_i & \geq & +1, \quad \text{for } x' = A_i \text{ and } D_{ii} = +1, \\ x'w & - & \gamma & - & y_i & \leq & -1, \quad \text{for } x' = A_i \text{ and } D_{ii} = -1. \end{array} \tag{5}$$

The 1-norm of the slack variable y is minimized with weight ν in (2). The quadratic term in (2), which is twice the reciprocal of the square of the 2-norm distance $\frac{2}{\|w\|_2}$ between the two bounding planes of (3) in the n-dimensional space of $w \in R^n$ for a *fixed* γ, maximizes that distance, often called the "margin". Figure 1 depicts the points represented by A, the bounding planes (3) with margin $\frac{2}{\|w\|_2}$, and the separating plane (4) which separates $A+$, the points represented by rows of A with $D_{ii} = +1$, from $A-$, the points represented by rows of A with $D_{ii} = -1$.

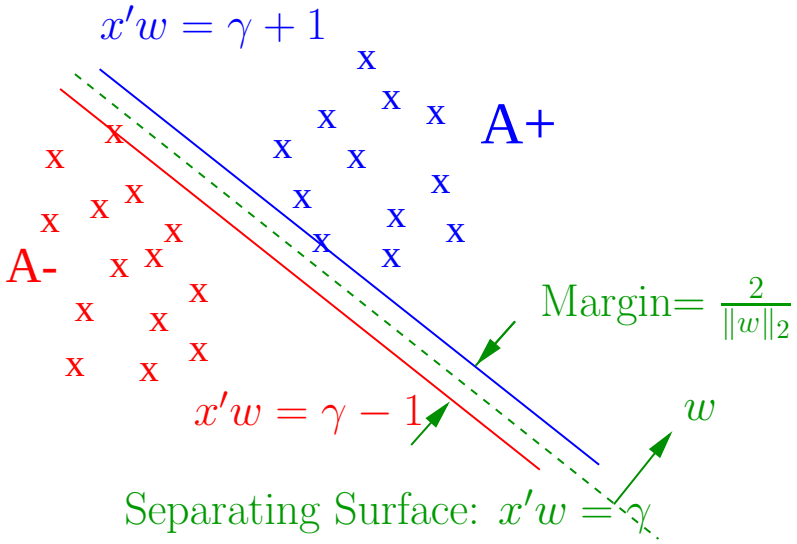

Figure 1: The bounding planes (3) with margin $\frac{2}{\|w\|_2}$, and the plane (4) separating $A+$, points represented by rows of A with $D_{ii} = +1$, from $A-$, points represented by rows of A with $D_{ii} = -1$.

In our smooth approach, the square of 2-norm of the slack variable y is minimized with weight $\frac{\nu}{2}$ instead of the 1-norm of y as in (2). In addition the distance between the planes (3) is measured in the $(n+1)$-dimensional space of $(w, \gamma) \in R^{n+1}$, that is $\frac{2}{\|(w,\gamma)\|_2}$. Measuring the margin in this $(n+1)$-dimensional space instead of R^n induces strong convexity and has little or no effect on the problem as was shown in [12]. Thus using twice the reciprocal squared of the margin instead, yields our modified SVM problem as follows:

$$\begin{array}{rll} \min\limits_{(w,\gamma,y)\in R^{n+1+m}} & \frac{\nu}{2}y'y + \frac{1}{2}(w'w + \gamma^2) & \\ \text{s.t.} & D(Aw - e\gamma) + y \geq e & \\ & y \geq 0. & \end{array} \tag{6}$$

At a solution of problem (6), y is given by

$$y = (e - D(Aw - e\gamma))_+, \tag{7}$$

where, as defined earlier, $(\cdot)_+$ replaces negative components of a vector by zeros. Thus, we can replace y in (6) by $(e - D(Aw - e\gamma))_+$ and convert the SVM (6) into an equivalent SVM which is an unconstrained optimization

problem as follows:

$$\min_{(w,\gamma)\in R^{n+1}} \quad \frac{\nu}{2}\|(e - D(Aw - e\gamma))_+\|_2^2 + \frac{1}{2}(w'w + \gamma^2). \tag{8}$$

This problem is a strongly convex minimization problem without any constraints. It is easy to show that it has a unique solution. However, the objective function in (8) is not twice differentiable which precludes the use of a fast Newton method. We thus apply the smoothing techniques of [3, 4] and replace x_+ by a very accurate smooth approximation that is given by $p(x, \alpha)$, the integral of the sigmoid function $\frac{1}{1+\varepsilon^{-\alpha x}}$ of neural networks [10], that is

$$p(x, \alpha) = x + \frac{1}{\alpha}\log(1 + \varepsilon^{-\alpha x}), \ \alpha > 0. \tag{9}$$

This p-function with a smoothing parameter α is used here to replace the plus function of (8) to obtain a smooth support vector machine **(SSVM)** :

$$\min_{(w,\gamma)\in R^{n+1}} \Phi_\alpha(w, \gamma) \ := \quad \min_{(w,\gamma)\in R^{n+1}} \frac{\nu}{2}\|p(e - D(Aw - e\gamma), \alpha)\|_2^2 + \frac{1}{2}(w'w + \gamma^2). \tag{10}$$

It was shown in [9] that the solution of problem (6) is obtained by solving problem (10) with α approaching infinity. The twice differentiable property of the objective function of (10) was utilized in [9] to obtain a globally and quadratically convergent algorithm for solving the smooth support vector machine (10). This was implemented computationally in [9] and here as follows. When we computed the Newton descent direction, we used the limit values of both the sigmoid function $\frac{1}{1+\varepsilon^{-\alpha x}}$ and the p-function (9) as the smoothing parameter α goes to infinity, that is the unit step function with value $\frac{1}{2}$ at zero and the plus function $(\cdot)_+$, respectively.

We briefly describe how the generalized support vector machine (GSVM) [11] generates a nonlinear separating surface by using a completely

arbitrary kernel. GSVM solves the following problem for a general kernel $K(A, A')$:

$$\begin{array}{rll} \min\limits_{(u,\gamma,y)\in R^{m+1+m}} & \nu e'y + f(u) & \\ \text{s.t.} & D(K(A, A')Du - e\gamma) + y \ \geq & e \\ & y \ \geq & 0. \end{array} \tag{11}$$

Here $f(u)$ is some convex function on R^m which suppresses the parameter u and ν is some positive number that weights the classification error $e'y$

versus the suppression of u. A solution of this problem, u and γ, leads to the nonlinear separating surface:

$$K(x', A')Du - \gamma = 0. \tag{12}$$

The linear formulation (2) is obtained if we let $K(A, A') = AA'$, $w = A'Du$ and $f(u) = \frac{1}{2}u'DAA'Du$. We now use a different classification objective which not only suppresses the parameter u but also suppresses γ in our nonlinear formulation:

$$\begin{array}{rll} \min\limits_{(u,\gamma,y)\in R^{m+1+m}} & \frac{\nu}{2}y'y + \frac{1}{2}(u'u + \gamma^2) & \\ \text{s.t.} & D(K(A, A')Du - e\gamma) + y \geq e & \\ & y \geq 0. & \end{array} \tag{13}$$

We repeat here the earlier arguments used to smooth the linear SVM to obtain the following SSVM with a nonlinear kernel $K(A, A')$:

$$\min_{(u,\gamma)\in R^{m+1}} \quad \frac{\nu}{2}\|p(e - D(K(A, A')Du - e\gamma), \alpha)\|_2^2 + \frac{1}{2}(u'u + \gamma^2), \tag{14}$$

where $K(A, A')$ is a kernel map from $R^{m\times n} \times R^{n\times m}$ to $R^{m\times m}$. We note that this problem, which is capable of generating highly nonlinear separating surfaces, still retains the strong convexity and differentiability properties for any arbitrary kernel. Hence a very fast Newton algorithm [9] applies to it as well.

With the help of the above tools we turn to the problems of clustering and classifying breast cancer patients.

3 Clustering Procedure: Determining Patient Groups

In this section we try to identify the group of patients who could benefit from chemotherapy by using k-Median clustering algorithm [2]. We will cluster the 253 breast cancer patients listed in the publicly available dataset [19] into three groups: Good, Intermediate and Poor, that will strongly reflect distinct patient survival times. We obtained our groupings using five cytological features (mean area, standard error of area, worst area, worst texture and worst perimeter) determined from a fine needle aspirate taken

during diagnosis using the XCYT diagnostic system [17, 20, 21, 13] and one pathology feature (tumor size) determined from the surgical removal of the tumor. These six features were the same as those used in [8] and arrived at using feature selection techniques [1]. One additional feature determined from surgery, lymph node status (number of metastasized lymph nodes) will be used only as a criterion in determining the groups but not as a feature in the final classification of patients into these groups. The reason for not including lymph node status in the final classification is that determining it necessitates a possibly debilitating procedure for the patient which consists of removing the lymph nodes from the armpit of the patient. We note that lymph node status, which varies between 0 and 30 in number, is often used as a guide whether to use or not to use chemotherapy. A similar role is played by tumor size, which varies between 0.4 and 10 centimeters.

We term patients who have received chemotherapy as *chemo-patients* and those who have not received chemotherapy as *nochemo-patients*. Intuitively, the group that we want to identify should consist of the chemo-patients who have a good survival rate and the nochemo-patients who have a poor survival rate. Thus if we can cluster the chemo-patients and the nochemo-patients into two groups, "good" and "poor" groups, respectively, and merge the "good" chemo-patients and the "poor" nochemo-patients, then we have the desired group of patients for which chemotherapy might prolong survival. Based on physicians' knowledge and experience, we use tumor size and lymph node status as criteria to define a "good condition" and a "poor condition" for breast cancer patients as follows:

- Good condition: Patients who have no metastasized lymph node and their tumor size < 2 centimeters.
- Poor condition: Patients either have more than 5 metastasized lymph nodes or their tumor size ≥ 4 centimeters.

We then compute the median centers of good condition patients and poor condition patients in the 6-feature space respectively. The median centers are used as the initial cluster centers in a k-Median clustering algorithm [2] because clustering results are highly dependent on the initial cluster centers and we want the final clusters to inherit the same good and poor conditions of the cluster centers. Next we use the k-Median algorithm to cluster the 113 nochemo-patients into two groups. The group that is clustered around

the median of the good condition patients is called NoChemoGood. The group that is clustered around the median of the poor condition patients is called NoChemoPoor. We then cluster the 140 chemo-patients using the same initial cluster centers that were used in no-chemo patients. This generates ChemoGood, the group of chemo-patients clustered around the median of the good condition patients, and ChemoPoor, the group of chemo-patients clustered around the median of the poor condition patients. We finally merge NoChemoPoor and ChemoGood together to obtain the Intermediate group. All the computations were run on the University of Wisconsin Computer Sciences Department Ironsides cluster. This cluster of four Sun Enterprise E6000 machines, each machine consisting of 16 UltraSPARC II 250MHz processors and 2 gigabytes of RAM, resulting in a total of 64 processors and 8 gigabytes of RAM.

We now summarize the above clustering procedure for generating our three groups in the following three steps:

Step 1: Compute the initial cluster centers:

- Compute the median of the good condition patients (Patients that do not have metastasized lymph node and their tumor size < 2 centimeters)
- Compute the median of the poor condition patients (Patients that either have more than 5 metastasized lymph nodes or their tumor size ≥ 4 centimeters)

Step 2: Use as initial cluster centers the medians of good condition patients and poor condition patients from Step1, in the k-Median clustering algorithm [2] to:

- Cluster the 113 nochemo-patients into:
  - NoChemoGood (clustered around the median of good condition patients)
  - NoChemoPoor (clustered around the median of poor condition patients)
- Cluster the 140 chemo-patients into:
  - ChemoGood (clustered around the median of good condition patients)

– ChemoPoor (clustered around the median of poor condition patients)

Step 3: Obtain the final three groups as follows:

- Good: NoChemoGood from Step 2
- Poor: ChemoPoor from Step 2
- Intermediate: NoChemoPoor and ChemoGood from Step 2

These three steps are depicted in the flow chart of Figure 2.

The three groups obtained in Step 3 above have very well separated survival curves as shown in Figure 3. In addition, the p-value of the logrank statistic [7] between any two groups is less than 0.0076. Thus, we tend to reject the null-hypothesis (survival curves are the same) and conclude that these survival curves are indeed significantly different from each other.

We note that a principal objective of this grouping is to generate the Intermediate group in Step 3 which contains 67 chemo-patients and 44 nochemo-patients. The chemo-patients have a better survival curve than the nochemo-patients in this group, as shown in Figure 4. This property is the reverse case to that of the entire 253 patients population as depicted in Figure 5, where the nochemo-patients have a better survival curve than the chemo-patients. To our knowledge such explicit benefit of chemotherapy over no-chemotherapy has not been quantified in the literature.

Another interesting property of the Intermediate group is the reversal of survival curves for the chemo-patients with lymph node positive and nochemo-patients with lymph node negative. That is, in the Intermediate group the subgroup with chemo-patients with lymph node positive has better survival than the subgroup with nochemo-patients with lymph node negative as shown in Figure 6. In contrast, in the overall population the reverse is true as shown in Figure 7. All above comparisons between the survival curves are significantly different based on the logrank statistic.

We note that the clustering procedure above cannot be utilized on a new patient since we want to exclude both chemotherapy (because it is unavailable) and lymph node status (because we want to forgo this risky procedure) in assigning a patient to a survival group. Therefore we turn now to classifying all 253 patients into three groups, Good, Intermediate and Poor, using a procedure consisting of three nonlinear support vector machine classifiers.

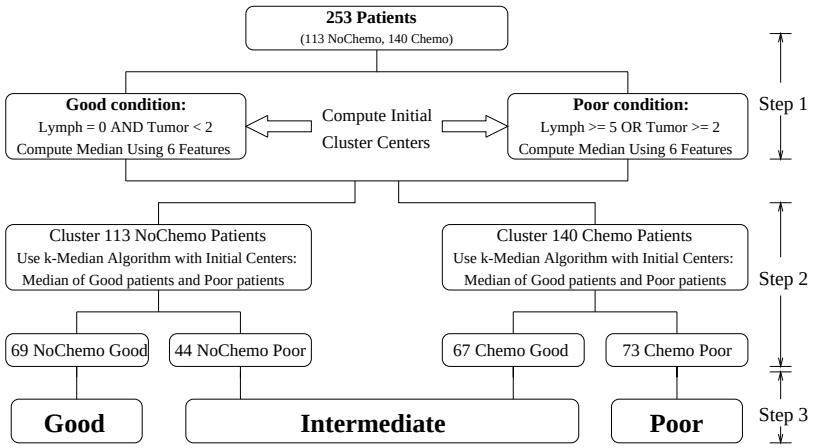

Figure 2: Clustering flow chart of 253 breast patients into three groups: Good, Intermediate and Poor

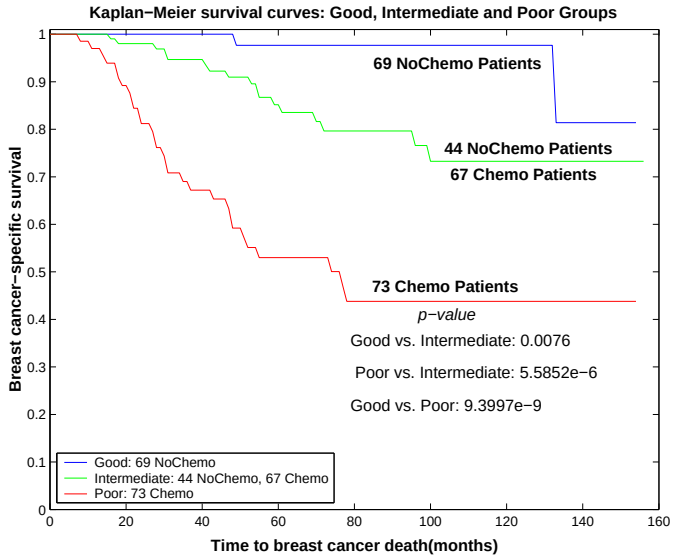

Figure 3: Kaplan-Meier survival curves for the Good, Intermediate and Poor groups.

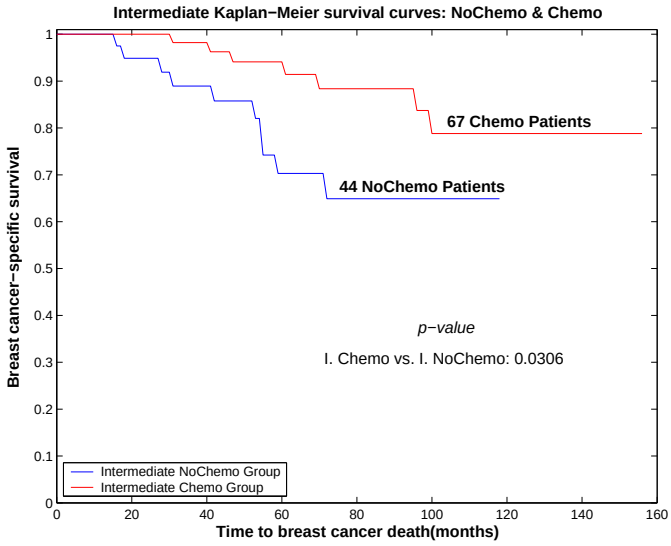

Figure 4: Kaplan-Meier survival curves for patients in the Intermediate group split into two groups: those who have had chemotherapy and those who have not.

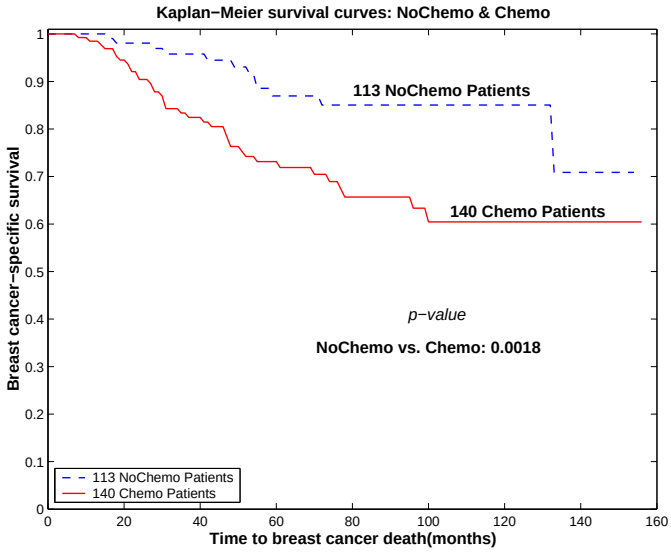

Figure 5: Kaplan-Meier survival curves for the **overall** patients split into two groups: those who have had chemotherapy and those who have not.

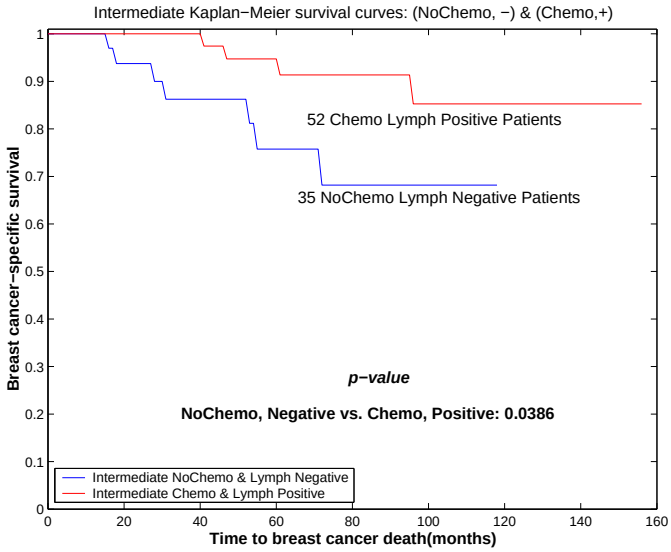

Figure 6: Kaplan-Meier survival curves for patients in the Intermediate group split into two groups: those with lymph node positive who have had chemotherapy and those lymph node negative who have not had chemotherapy.

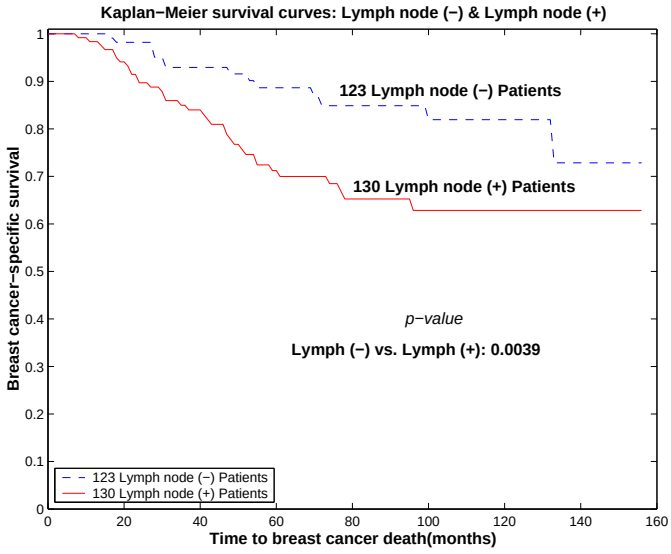

Figure 7: Kaplan-Meier survival curves for the **overall** patients split into two groups: those with lymph node negative and those with lymph node positive

4 A Support Vector Machine Prognostic Classification Procedure

In this section, we describe a procedure for classifying breast cancer patients into the three prognostic groups: Good (NoChemoGood), Intermediate (ChemoGood and NoChemoPoor) and Poor (ChemoPoor), generated by clustering in the previous section (*i.e.*, we assign each patient a class label according to our clustering results) based only on the 6 available features at time of diagnosis and selected by a support vector machine. This is a *multicategory* classification problem that cannot be solved by a single SVM classifier. Our proposed classification procedure described below is very similar to DAGSVM, the directed acyclic graph SVM procedure of [15]. The classification process utilizes three classifiers (Good *vs.* Poor; Good *vs.* ChemoGood; and NoChemoPoor *vs.* Poor) described in the following steps. We use the 5 cytological features (mean area, standard error of the area, worst area, worst texture and worst perimeter) and tumor size in our procedure. In order to exclude both lymph node status and the chemotherapy indicators in assigning a patient to a survival group, we define the lymph node index LI and the chemo index CI in our classification procedure to simulate the patient's lymph node status and the chemotherapy indicators respectively. Both LI and CI are generated by a nonlinear SSVM classifier and depend only on the 6 selected features. All classification was carried out using the nonlinear SSVM with the Gaussian kernel (1) that was described in Section 2 and implemented by using standard native MATLAB commands [14]. The whole procedure includes seven nonlinear SSVM classifiers (three for solving the multicategory classification problem and four for generating the lymph node index and the chemo index). The largest case of these seven classifiers is generated by solving a nonlinear SSVM in R^{139} real space and each classifier can be generated less than 3 CPU seconds. We outline our classification procedure now.

Step 1: Separate the Good group from the Poor group by a nonlinear SSVM with a Gaussian kernel. We call this nonlinear classifier **SVM1**. **SVM1** achieves 92% tenfold test set correctness for this classification.

Step 2: Label patients Good1 and Poor1 as follows:

- Good1: ChemoGood and NoChemoGood (obtained in the clustering Step 2 of Section 3) consisting of 136 patients

- Poor1: ChemoPoor and NoChemoPoor (obtained in the clustering Step 2 of Section 3) consisting of 117 patients

Step 3: Generate a lymph node index LI(x) for each of the two groups, Good1 and Poor1 above, by separating within each group lymph node positive patients from lymph node negative patients using the 6 features specified above:

- LI(x):=$K(x', A^{1'})D^1u^1 - \gamma^1$ for x in Good1
  - LI(x) > 0 surrogate for lymph node positive
  - LI(x) ≤ 0 surrogate for lymph node negative
- LI(x)=$K(x', A^{2'})D^2u^2 - \gamma^2$ for x in Poor1
  - LI(x) > 0 surrogate for lymph node positive
  - LI(x) ≤ 0 surrogate for lymph node negative

Here and in Step 4 below, $K(\cdot, A^{i'})$ is the Gaussian kernel (1), the matrix $A^1 \in R^{136\times 6}$ represents the patients in Good1 and $A^2 \in R^{117\times 6}$ those in Poor1, while the diagonal matrix D^1 of ± 1 labels lymph node positive and lymph node negative patients respectively in Good1. D^2 does the same for patients in Poor1. Furthermore, (u^i, γ^i), $i = 1, \ldots, 4$, is a solution of (14) with $A = A^i$ and $D = D^i$.

Step 4: Generate a chemo index CI(x) for each group, Good1 and Poor1 above by separating within each group chemo-patients from nochemo-patients using the same 6 features:

- CI(x)=$K(x', A^{1'})D^3u^3 - \gamma^3$ for x in Good1
- CI(x)=$K(x', A^{2'})D^4u^4 - \gamma^4$ for x in Poor1

Here, the diagonal matrix D^3 of ± 1 labels patients with and without chemotherapy respectively in Good1 and D^4 does the same for patients in Poor1.

Step 5: Separate NoChemoGood from ChemoGood in Good1 and NoChemoPoor from ChemoPoor in Poor1 respectively. These two classifiers are obtained by using a Gaussian kernel on the 6 original features combined with a linear kernel on the lymph node index LI and the chemo index CI (using 8 features in all, two of which LI and CI dependent on the

original 6 features) on each of the sets Good1 and Poor1. The nonlinear classifier that is used for classifying Good1 is called **SVM2** and the classifier used for classifying Poor1 is called **SVM3**.

Step 6: Obtain the final three groups:

- Good = The NoChemoGood separated from ChemoGood within Good1
- Poor = The ChemoPoor separated from the NoChemoPoor within Poor1
- Intermediate = ChemoGood within Good1 AND NoChemoPoor within Poor1

We summarize the classification procedure above in the flow chart depicted in Figure 8.

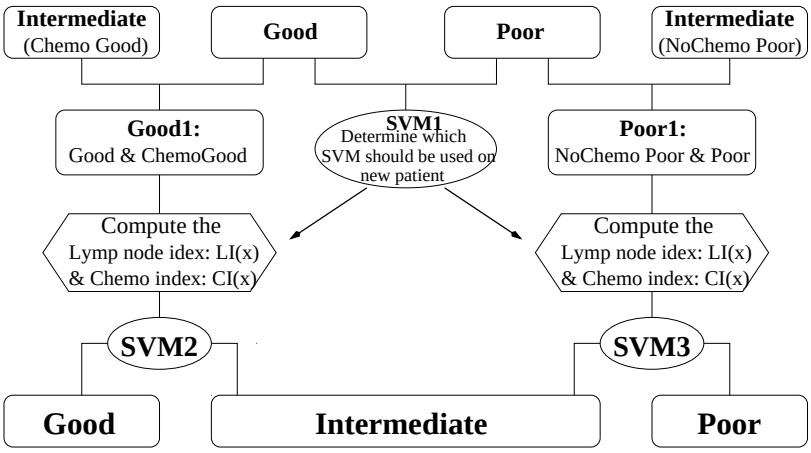

Figure 8: Flow chart for classifying 253 breast cancer patients into the three groups using three SVMs, Good, Intermediate and Poor groups that were generated in Section 3.

Having separated the above three groups, with tenfold test set correctness of 82.7%, we prescribe the following procedure for classifying a new patient into one of the three groups: Good, Intermediate and Poor. This leads to the following prognostic procedure.

Prognostic Procedure 4.1: Given 6 features: 5 cytological features (mean area, standard error of the area, worst area, worst texture and worst perimeter) and tumor size for a new patient:

(i) Label the patient as Good1 or Poor1 as defined in Step 2 by using the **SVM1** classifier of Step 1.

(ii) Generate a lymph node index LI(x) and a chemo index CI(x) for the patient as defined in Steps 3 and 4 which depends on the 6 features only.

(iii) Classify the patient into Good, Poor or Intermediate by using one of the two classifiers, of Step 5 above, depending on whether the patient has been classified into Good1 or Poor1 in (i) above.

We summarize the overall prognostic procedure above in the flow chart depicted in Figure 9.

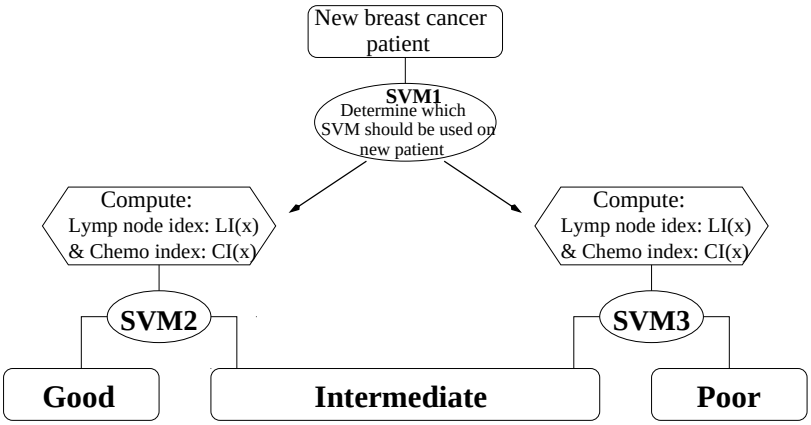

Figure 9: Flow chart for the overall classification procedure for a new breast cancer patient into one of the three groups: Good, Intermediate and Poor.

This procedure works for any new patient without knowing whether chemotherapy will be given to this patient, nor do we want to make the assumption that lymph node status is automatically available for that patient.

5 Conclusion

We have obtained a classification, with 82.7% test set correctness, of a publicly available 253 breast cancer patient dataset into three survival categories: Good, Poor and Intermediate. The survival curve for each group is very distinct from the others with the following additional properties:

- The Good group patients are all without chemotherapy.
- The Poor group patients are all with chemotherapy.
- The patients in the Intermediate group with chemotherapy have better survival than those in the same group without chemotherapy, which is the reverse that for the total population.

Based on the prognostic procedure we associate with the patient one of three survival, Good, Intermediate, and Poor, in Figure 3 with the corresponding longevity. These curves suggest that:

1. Good group patients should not receive chemotherapy.
2. Intermediate group patients should receive chemotherapy based on the two survival curves Figure 4 and Figure 5.

We believe that these are novel findings which will hopefully help doctors and patients in assessing post-operative longevity.

Acknowledgments

The research described in this Data Mining Institute Report 01-03, March 2001, was supported by National Science Foundation Grants CCR-9729842 and CDA-9623632, by Air Force Office of Scientific Research Grant F49620-00-1-0085 and by the Microsoft Corporation.

References

[1] P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In J. Shavlik, editor, *Machine Learning Proceedings of the Fifteenth International Conference(ICML*

'98), pages 82–90, San Francisco, California, 1998. Morgan Kaufmann. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-03.ps.

[2] P. S. Bradley, O. L. Mangasarian, and W. N. Street. Clustering via concave minimization. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems -9-*, pages 368–374, Cambridge, MA, 1997. MIT Press. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/96-03.ps.

[3] Chunhui Chen and O. L. Mangasarian. Smoothing methods for convex inequalities and linear complementarity problems. *Mathematical Programming*, 71(1):51–69, 1995.

[4] Chunhui Chen and O. L. Mangasarian. A class of smoothing functions for nonlinear and mixed complementarity problems. *Computational Optimization and Applications*, 5(2):97–138, 1996.

[5] V. Cherkassky and F. Mulier. *Learning from Data - Concepts, Theory and Methods*. John Wiley & Sons, New York, 1998.

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

[7] David G. Kleinbaum. *Survival Analysis*. Springer-Verlag, New York, 1996.

[8] Y.-J. Lee, O. L. Mangasarian, and W. H. Wolberg. Breast cancer survival and chemotherapy: a support vector machine analysis. Technical Report 99-10, Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, December 1999. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, American Mathematical Society, Volume 55, 2000, 1-10. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/99-10.ps.

[9] Yuh-Jye Lee and O. L. Mangasarian. SSVM: A smooth support vector machine. *Computational Optimization and Applications*, 20:5–22, 2001. Data Mining Institute, University of Wisconsin, Technical Report 99-03. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/99-03.ps.

[10] O. L. Mangasarian. Mathematical programming in neural networks. *ORSA Journal on Computing*, 5(4):349–360, 1993.

[11] O. L. Mangasarian. Generalized support vector machines. In A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 135–146, Cambridge, MA, 2000. MIT Press. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-14.ps.

[12] O. L. Mangasarian and D. R. Musicant. Successive overrelaxation for support vector machines. *IEEE Transactions on Neural Networks*, 10:1032–1037, 1999. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/98-18.ps.

[13] O. L. Mangasarian, W. N. Street, and W. H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43(4):570–577, July-August 1995.

[14] MATLAB. *User's Guide*. The MathWorks, Inc., Natick, MA 01760, 1994-2001. http://www.mathworks.com.

[15] J. C. Platt, N. Cristianini, and J. Shawe-Taylor. Large margin dags for multiclass classification. *Advances in Neural Information Processing Systems (NIPS2000)*, 12:547–553, 2000.

[16] W. N. Street, O. L. Mangasarian, and W. H. Wolberg. An inductive learning approach to prognostic prediction. In Armand Prieditis and Stuart Russell, editors, *Machine Learning: Proceedings of the Twelfth International Conference*, pages 522–530, San Francisco, 1995. Morgan Kaufmann.

[17] W. N. Street, W. H. Wolberg, and O. L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical Image Processing and Biomedical Visualization*, volume 1905, pages 861–870, San Jose, California, 1993. SPIE–The International Society for Optical Engineering.

[18] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, second edition, 2000.

[19] W. H. Wolberg, Y.-J. Lee, and O. L. Mangasarian. WPBCC: Wisconsin Prognostic Breast Cancer Chemotherapy Database.

Computer Sciences Department, University of Wisconsin, Madison, ftp://ftp.cs.wisc.edu/math-prog/cpo-dataset/machine-learn/cancer/WPBCC/, 1999.

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

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