Breast Cancer Identification: KDD CUP Winner's Report

Claudia Perlich, Prem Melville, Yan Liu,
Grzegorz Świrszcz, Richard Lawrence
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
{perlich,pmelvil,liuya}@us.ibm.com
{swirszcz,ricklawr}@us.ibm.com

Saharon Rosset
Raymond and Beverly Sackler School of
Mathematical Sciences
Tel Aviv University,
Israel
saharon@post.tau.ac.il

ABSTRACT

We describe the ideas and methodologies that we developed in addressing the KDD Cup 2008 on early breast cancer detection, and discuss how they contributed to our success. The most important components of our solution were 1) the identification of predictive information in the patient identifier, 2) a linear SVM on the 117 provided features, and 3) a heuristic post-processing approach to optimize the evaluation criteria.

1. TASK AND DATA DESCRIPTION

The KDD Cup 2008 was organized by Siemens medical solutions and consisted of two prediction tasks in breast cancer detection from images. The organizers provided data from 1712 patients for training; of these 118 had cancer. Siemens uses proprietary software to identify in each image (two views for each breast) suspect locations (called *candidates*), that are described by their coordinates and 117 features. No explanation of the features was given. Overall the training set includes 102,294 candidates, 623 of which are positive. A second dataset with similar properties was used as the test set for competition evaluation. The two modeling tasks were:

Task 1: Rank the candidates by the likelihood of being cancerous in decreasing order. The evaluation criterion for this task was an area under the FROC curve, which measures how many of the actual *patients* with cancer are identified while limiting the number of *candidate* false alarms to a range between 0.2 and 0.3 per image. This was meant to reflect realistic requirements when the prediction model is used as an actual decision support tool for radiologists.

Task 2: Suggest a maximal list of patients who are surely healthy. In this task, including any patient with cancer in the list will disqualify the entry. This was meant to be appropriate for a scenario where the model is used to save the radiologist work by ruling out patients who are *definitely healthy*, and thus the model was required to have *no false negatives*.

Several aspects of the data and the tasks made this competition interesting, including:

- The presence of leakage, whereby patient IDs turned carry predictive information about a patient's likelihood to have cancer. We discuss this competition's leakage and other examples in Section 2.
- Unique data properties, including the presence of extreme outliers and the combination of the features with neighborhood-based information from the location of candidates. These properties and some of our efforts in alleviating and using them, are discussed in Section 3.
- The unique FROC score, which treats patients as positive examples, but candidates as negative examples. This clearly has implications on the way in which models should rank candidates, preferentially combining candidates from different patients over many good candidates from the same patient. We address this in the context of post-processing schemes for model scores in Section 4.

We present our final submission briefly in Section 5.

2. LEAKAGE IN PATIENT ID

Leakage can be defined as the introduction of predictive information about the target by the data generation, collection, and preparation process. Such information leakage - while potentially highly predictive out-of-sample *within* the study - leads to limited generalization and model applicability, and to overestimation of the predictive performance. Two of the most common causes for leakage are:

1. Combination of data from multiple sources and/or multiple time points, followed by a failure to completely anonymize the data and hide the different sources.
2. Accidental creation of artificial dependencies and additional information while preparing the data for the competition or proof-of-concept.

This year's KDD Cup data suffered from leakage that was probably due to the first cause. The patient IDs in the competition data carried significant information towards identifying patients with malignant candidates. This is best illustrated through a discretization of the patient ID range, as demonstrated in Figure 1. The patient IDs are naturally divided into three disjoint bins: between 0 and 20,000 (254 patients; 36% malignant); between 100,000 and 500,000 (414 patients; 1% malignant); and above 4,000,000 (1044 patients, of them 1.7% malignant). We can further observe that all 18 afflicted patients in the last bin have patient IDs in the range 4,000,000 to 4,870,000, and there are only 3 healthy patients in this range. This gives us a four-bin division of the data with great power to identify sick patients. This binning and its correlation with the patient's health

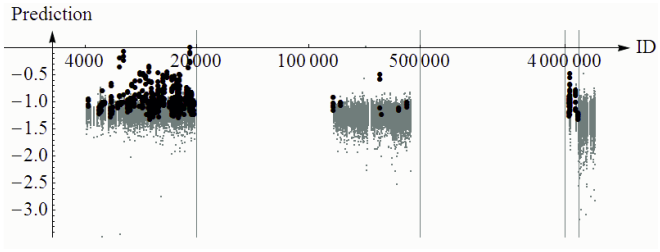

Figure 1: Distribution of malignant (black) and benign (gray) candidates depending on patient ID on the X-axis in log scale. The Y-axes is the score of a linear SVM model on the 117 features. Vertical lines show the boundaries of the identified ID bins.

generalized to the test data. Our hypothesis is that this leakage reflects the compilation of the competition data from different medical institutions and maybe different equipment, where the identity of the source is reflected in the ID range and is highly informative of the patient's outcome. For example, one source might be a preventive care institution with only very low base rate of malignant patients and another could be a treatment-oriented institution with much higher cancer prevalence[1].

While it is clear that such leakage does not represent a useful pattern for real application, we consider its discovery and analysis an integral and important part of successful data analysis. Furthermore, we suggest that the problem in this dataset may actually run deeper and that all participants unknowingly benefited somewhat from a leakage problem and all reported performances are likely to be inflated.

If the predictiveness of the identifiers is caused by the combination of data from different sources, there may be additional implicit leakages due to differences in data collection settings (e.g., machine calibration). This would still be present even if the patient IDs had been removed. We test this hypothesis with the following experiment: If such a leakage exists (say the average grayscale is slightly different), it should be possible to predict the data source (i.e., one of the four identifier bins) from negative candidates only. We cannot include positives because we already know that the cancer prevalence is correlated with the bins. Our analysis shows that both group 1 (ID below 20000) and group 4 (ID above 4,870,000) are easily identified by a logistic model from the 117 provided features with AUCs of 0.86 and 0.75 respectively.

Given this result we feel confident to conclude that any reasonable model can infer the patient group to some extent from the 117 variables and thereby implicitly the cancer prevalence in that patient population. So all models built on this data set are likely to overestimate the true predictive performance of cancer detection when applied to an entirely different population.

More generally, experience has shown that leakages occur in many modeling competitions, including KDD-Cup 2007 [4], where the organizers' preparation of the data for one task exposed some information about the response for the other task; the INFORMS Data Mining Contest in 2008, were it was possible to identify the partial removal of diagnosis codes used for the target construction; and KDD-Cup 2000 [1], where internal testing patterns that were left in the data by the organizers supplied a significant boost to those who were able to identify them.

Exploratory data analysis seems to have become something of a lost art in the KDD community. In proper exploratory analysis, the modeler carefully examines the data with little preconception about what it contains, and allows patterns and phenomena to present themselves, only then analyzing them and questioning their origin and validity. We hope that our discovery of this leakage can serve as a reminder of the value of open-minded exploratory analysis.

[1]The organizers later explained that in order to increase the number of positive examples, the dataset was comprised of examples from different time periods.

3. MODELING

3.1 Incorporating IDs

Given the obvious predictive value of the patient ID we incorporated this information as a categorical variable for the classification models with 4 possible bin numbers {1,2,3,4}. We also explored building 4 separate models, but this did not yield better results, presumably because for some of them the number of training examples is rather small.

3.2 Classification

In order to investigate the generalization performance of different methods, we created a stratified 50% training and test split by patient. We ensured that exactly half of the positive patients were assigned to each split. All results presented in Table 1 are based on our internal test set[2].

We explored the use of various learning algorithms for the underlying candidate classification problem including Neural Networks, Logistic regression and several SVM variants using the *SVMPerf* package [2]. Ultimately, linear models (logistic regression or linear SVMs) yielded the most promising results. In this section we explore various directions for improving the initial FROC results from the linear SVM of 0.0834 without and 0.0882 with a patient bin variable.

Kernel selection: We compared linear SVMs, RBF kernels and polynomial kernels of degree 2 and 3. We found that linear kernels performed best, and have the added advantage of being extremely fast compared to the other approaches. The RBF kernels took the longest time to run and had a dismal performance of 0.0229. Given these results, we adopted linear SVMs for all following experiments.

Loss function: Most work in the use of SVMs has focused on minimizing the error rate or zero-one loss function. In recent work, Joachims [2] presented efficient ways to train SVMs to maximize alternative multivariate performance measures, such as the area under the ROC curve (AUC). Given that the evaluation metric for Task 1 is related to AUC, we trained an SVM to maximize AUC. We also compared maximizing Precision and Recall at k, which is the Precision/Recall of a classifier that predicts exactly k instances as positive. In particular, given p positive instances in the training set, we used $k = p/2$ and $k = 2p$ for Precision and Recall respectively. Since AUC is most closely related to the FROC metric, we find that maximizing it performs the best, improving the FROC from 0.0882 to 0.0893 but probably not significantly.

[2]The labeles of the true test set were never published.

Approach	FROC
Linear SVM without ID bin	0.0834
Linear SVM with ID bin	0.0882
Linear SVM without Outliers	0.0858
Polynomial kernel SVM, d=2	0.0803
Polynomial kernel SVM, d=3	0.0774
RBF kernel SVM	0.0229
Linear SVM maximizing AUC	0.0893
Linear SVM maximizing Precision	0.0869
Linear SVM maximizing Recall	0.0865
Bagging linear SVM with c=20	0.0900
Bagging linear SVM c=500	0.0873
Constrained Logistic Regression	0.0793
Bagged Linear SVM with Post Processing	0.0930

Table 1: Comparing FROC of different approaches on our 50% test set.

Regularization and Bagging: As observed by Valentini and Dietterich [5], bagging can significantly improve classification accuracy over a single SVM. Since bagging is a variance-reduction technique, they propose applying bagging to SVMs with low bias and high variance. In particular, for linear SVMs, they show that decreasing bias by increasing the regularization parameter and then applying bagging is very effective. We test the effectiveness of bagging in our setting for maximizing FROC. We applied 10 iterations of bagging linear SVMs with the regularization parameter c set to 20 and 500. We observed that with $c = 20$, bagging does improve the FROC to 0.090. For $c = 500$, however the FROC drops to 0.0873, so the accuracy results do not seem to carry over in our setting.

3.3 Outlier Treatment

The majority of the 117 features exhibit distributions with heavy tails and at times significant skew. We explore cutting outliers to improve our models estimation and to avoid extreme predictions caused by extreme x values on the submission set. We calculated an upper and lower bound for each of the 117 features based on the inter-quartile range. We define Q_k^{25} as the 25th and Q_k^{75} as the 75th percentile of feature k. We replaced each value $x_k < Q_k^{25} - 3*(Q_k^{75} - Q_k^{25})$ with the lower bound $Q_k^{25} - 3*(Q_k^{75} - Q_k^{25})$ and $x_k > Q_k^{75} + 3*(Q_k^{75} - Q_k^{25})$ with the upper bound $Q_k^{75} + 3*(Q_k^{75} - Q_k^{25})$. This adjustment affected a non-negligible percentage of candidates and reduced the linear SVM performance to 0.0858 and so we kept the extreme values unaltered.

3.4 Additional Features

Candidate Location: We considered that the location of a candidate relative to nipple or breast boundary might be indicative of malignant candidates. Figure 2 shows the normalized and aligned locations of all candidates. It turned out that even sophisticated features derived from the location such as high-degree polynomials, functions of the distance from the deduced boundaries etc. yielded no improvement. This indicates that either there are no "high risk" areas in the breast, or that the 117 original features already contain the relevant location information of the candidate.

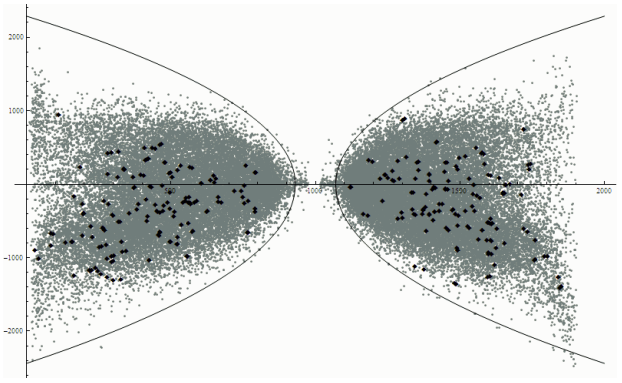

Figure 2: Distribution of location of malignant (black diamond) and benign (gray dot).

Number of candidates per patient: Another direction that could potentially lead to additional informative features is the number of candidates belonging to a patient. Once again, all the attempts to construct features using this quantity lead either to deterioration or non–substantial improvement of the performance on the test data.

3.5 Modeling Multiple Views

As suggested in one of the hints provided by the organizers, a cancerous lesion should be visible in both views v_1 and v_2 of a breast (although in extremely rare cases some lesions may only be visible in one view). It might therefore be possible to formulate meaningful constraints and exploit correlations in the classification decisions for candidates from the same region of a breast.

Constraint Formulation: The data contains the coordinates of each candidate as well as the coordinates of the nipple in each image. Since the different views have different directions, we cannot compare the coordinates directly. Rather we identify pairs of candidates in terms of similar Euclidean distance to the nipple in both views for each breast. We define a match as a pair of candidates from different views with the difference of the Euclidean distances less than a threshold. For a threshold of 20 this translated into 29139 matched pairs $\{(x_{k,v_1}, x_{k,v_2})\}_{k\in C}$, where C denotes the set of (29139) indexes. We would like candidates belonging to a pair to have similar predicted labels, i.e. $f(x_{k,v_1}) = f(x_{k,v_2})$. We shall incorporate this condition into an optimization formula as a penalty term.

Pairwise Constraint Kernel Logistic Regression: We use pairwise kernel logistic regression, which is able to plug in additional pairwise constraints together with labeled data to model the decision boundary directly [6]. Suppose we have a set of training examples $\{(x_i, y_i)\}$, and a set of pairs $\{(x_{k,v_1}, x_{k,v_2})\}_{k\in C}$ constructed from both labeled and unlabeled data. To make the optimization problem feasible to solve, we define a convex loss function via the logit loss as follows:

$$\mathcal{O}(f) = \frac{1}{m}\sum_{i=1}^{m}\log(1+e^{-y_i f(x_i)}) + \lambda\Omega(\|f\|_{\mathcal{H}}) +$$

$$\frac{\mu}{n}\sum_{k\in\mathcal{C}}\log(1+e^{f(x_{k,v_1})-f(x_{k,v_2})}) + \log(1+e^{f(x_{k,v_2})-f(x_{k,v_1})}),$$

where the first term is the loss on labeled training examples, the second is the regularizer and third term is the loss associated with the difference between the predicted labels of the example pairs. The pairwise constraint coefficient μ is set to 1. For simplicity, we define f as a linear classifier, i.e. $f(x) = w^T x$. Since the optimization function is convex, a gradient search algorithm can guarantee the finding of the global optimum. It is easy to derive the parameter estimation method using the interior-reflective Newton method, and we omit the detailed discussion. The constrained logistic regression unfortunately only yielded a FROC of 0.079.

4. POSTPROCESSING

Task 1 uses the predictions from the classification model to order the candidates and computes the FROC metric based on this ordering. As observed in Section 3 attempting to maximize AUC improves ranking of candidates, which in turn often improves FROC. However, optimizing AUC is not guaranteed to optimize FROC. The AUC measures the area under the curve of true positive rate versus false positive rate, whereas the Y-axis of a FROC curve is the true positive rate at a patient level. Contrary to AUC, a higher true positive rate at a candidate-level does not improve FROC unless the positive candidates are from different patients. For instance, it is better to have 2 correctly identified candidates from different patients, instead of 5 correctly identified candidates from the same. So it might be possible to reorder candidates such that a larger variety of patients are represented at the top of the list[3].

In order to do this, we create a pool of the top n candidates, as ordered by our model. We then select the candidates with the highest scores for each patient in this pool, and move these to the top of our list. We repeat this process iteratively with the remaining candidates in our pool until we have exhausted all candidates.

We only do this for the top n candidates, since the FROC metric is based only on the area under the curve for a small range of false alarm rates at the beginning of the curve. We leave the ordering of the remaining candidates untouched. The only parameter this post-processing procedure requires is the choice of n for the number of top-ranked candidates we want to re-order. We can select this parameter based on an estimate of the maximum number of candidates that have to be classified as positive before we hit the upper bound of the false alarm rate used in the FROC metric. For the specific FROC metric used to evaluate Task 1, it can be shown that the optimum value for n is the *number of positive candidates* $+ 1.2 \times$ *number of patients*.

Since the true number of positive candidates in the test set is not known, we estimate this from the positive rate in the training set. For our train-test split we used $n = 1500$, and report the results before and after post-processing in Table 1. This re-ordering of model scores had a significant impact on the resulting FROC, increasing the FROC of the bagged linear SVM model score from 0.09 to 0.093.

5. SUBMISSIONS AND RESULTS

[3]We have a theoretical result of an optimal reordering algorithm under the assumption that the model predictions are the correct probabilities of candidate malignancy. However, our attempts to calibrate the predictions failed to reach sufficient quality to take advantage of this optimality result.

Task 1: For our final submission we used the bagged linear SVM with the ID features, maximizing zero-one loss, $c = 20$ and heuristic post processing. This approach scored the winning result of 0.0933 on the test set.

Task 2: While the ID leakage had some influence on our Task 1 model, it was clearly essential for our Task 2 submission. It also explains the huge gap between our submission and the next best result. We investigated different models with and without ID features after ranking increasingly by the maximum candidate score of each patient. Models without ID typically produce the first false positive patient within about 200 patients of the training set (interestingly they are the positive patients from bin 3 just at the boundary between the two bins as seen in Figure 1). Conversely, we were fairly confident that bin 4 alone (more than 1000 on the training and 970 on the submission set) had no positive patients at all. Models with ID features rank all patients of bin 4 first and next some patients from bin 2. The first false negatives occur typically around 1100. On this task - a logistic regression model performs slightly better than the linear SVM models due to the high sensitivity of likelihood to extreme errors. We finally submitted the 1020 first ranked patients from a logistic model that included the ID features in addition to the original 117 provided features.

6. REFERENCES

[1] A. Inger, N. Vatnik, S. Rosset, and E. Neumann. KDD-CUP 2000: Question 1 winner's report. *SIGKDD Explorations*, 2000.

[2] T. Joachims. A support vector method for multivariate performance measures. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 377–384, New York, NY, USA, 2005. ACM.

[3] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*. MIT Press, 1998.

[4] S. Rosset, C. Perlich, and Y. Liu. Making the most of your data: Kdd cup 2007 "how many ratings" winner's report. *SIGKDD Explorations*, 2007.

[5] G. Valentini and T. G. Dietterich. Low bias bagged support vector machines. In *Proceedings of 20th International Conference on Machine Learning (ICML-2003)*, pages 752–759, Washington, DC, 2003.

[6] R. Yan, J. Zhang, J. Yang, and A. Hauptmann. A discriminative learning framework with pairwise constraints for video object classification. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'04)*, 2004.