Improving medical predictive models via Likelihood Gamble Pricing

Glenn Fung GLENN.FUNG@SIEMENS.COM
Harald Steck HARALD.STECK@SIEMENS.COM
Shipeng Yu SHIPENG.YU@SIEMENS.COM
Phan Giang PHAN.GIANG@SIEMENS.COM
IKM CKS, Siemens Medical Solutions, USA

Abstract

A combination of radiotherapy and chemotherapy, is often the treatment of choice for cancer patients. Recent developments in the treatment of patients have lead to improved survival. However, traditionally used clinical variables have poor accuracy for the prediction of survival and radiation treatment side effects. The objective of this work is to develop and validate improved predictive model for a large group of non-small cell lung cancer (NSCLC) patients and a group of rectal cancer patients. The main goal is to predict survival for both groups of patients and radiation induced side-effects for the NSCLC patients. Given sufficiently accurate predictions of these models, they can then be used to optimize the treatment of each individual patient, which is the goal of personalized medicine. Our improved predictive models are obtained by using the recently proposed Likelihood gamble pricing (LGP), which is a decision-theoretic approach to statistical inference that marries the likelihood principle of statistics with Von Neumann-Morgensterns axiomatic approach to decision making. The regularization induced by the LPG approach produces better probabilistic predictions than both the unregularized and the regularized (by the standard 2-norm regularization) widely used logistic regression approaches.

1. Introduction

Radiotherapy, combined with chemotherapy, is the treatment of choice for a large group cancer patients. For lung cancer patients, radiotherapy is not restricted to patients with mediastinal lymph node metastasis, but is also indicated for patients who are inoperable because of their physical condition. Improved radiotherapy treatment techniques allowed an increase of the radiation dose delivered to the tumor, while keeping the occurrence of treatment-induced side-effects, like Lung toxicity (pneumonitis) or dysphagia (esophagitis), to a minimum. Moreover, it has been found that combining radiotherapy with chemotherapy can improve outcomes even further. Several effective chemo radiation schemes are being applied. These developments have led to improved outcome in terms of survival time while minimizing the occurrence of treatment-related side-effects.

Given the multitude of chemo radiation schemes, it would be desirable to choose the most effective one for each individual patient. The relationship between patient/tumor characteristics, the applied treatment regime and the outcome is not well understood in the medical field to date. For this reason, we aim to learn predictive models that are able to estimate the outcome (survival and side-effects) for an individual patient under the various treatment options. Given sufficiently good prediction accuracy, these models can then be used to optimize the treatment of each individual patient, which is the goal of personalized medicine. These models can thus assist the medical doctor with decision support at the point of care.

Another example of personalized medicine is to predict tumor response after chemo-radiotherapy for locally advanced rectal cancer. This is important in individualizing treatment strategies, since patients with a pathologic complete response (pCR) after therapy, i.e., with no evidence of viable tumor on pathologic analysis, would need less invasive surgery or another

Appearing in the Proceedings of the ICML/UAI/COLT 2008 Workshop on Machine Learning for Health-Care Applications, Helsinki, Finland, 2008. Copyright 2008 by the author(s)/owner(s).

radiotherapy strategy instead of resection.

While several work of learning such models from existing data have been conducted, an accurate estimation of the survival probability to offer assistance for treatment decision-making for an individual patient is currently not available. This motivates the need for the creation of improved robust predictive models that take into account the available information about the patient prior to (and then during) treatment: tumor characteristics, clinical and demographic data as well as information on treatment (alternatives).

It is important to note that for survival models, the proportional hazards model (Cox regression model) is often the method of choice (Cox & Oakes, 1984). The proportional hazards model is a semi-parameter method that relates the time of an event (death or failure) to certain set of given predictive variables. However since we are interested in a model that also could predict probabilities of side effects (Lung toxicity, dysphagia, etc), we are assuming for this work that the survival model predicts the probability of survival at two years (binary classification problem).

We are specially interested in predicting accurate calibrated probabilities of survival or side-effects. In recent years, the interest for tools that could assists doctors in the clinical practice for diagnosis and prognosis is on the rise. Some of this methods and tools are based on probabilistic models since they results are relatively easy to understand when expressed as probabilities, hence the increasing interest in improving accuracies probabilities predictions that are in concordance with what a doctor may expect given the evidence (Szolovits & Pauker, 1990). There have been numerous papers that aim to convert predictive models outputs to probabilities and calibrated probabilities (Drish, 2001; Platt, 2000; Zadrozny & Elkan, 2001; Zadrozny & Elkan, 2002) but as far as we know few or none work has been applied to probabilistic calibrations in medical settings.

In several survival and side effects models, we find that the Likelihood Gamble Pricing (LGP) approach (Giang, 2006; Giang & Shenoy, 2005; Giang & Shenoy, 2002) provides excellent generalization, yielding more accurate and more robust predictions than other standard machine learning approaches.

In the experimental Section, we validate several prediction models for 2-year survival and side-effects (Lung toxicity and dysphagia) of non-small cell lung cancer (NSCLC) patients, treated with (chemo) radiotherapy, taking into account all available and established prognostic factors. The results illustrate the main properties of the LGP approach and of our computationally efficient approximation, and show that it achieves superior generalization and higher prediction accuracy than competing approaches on these medical data sets.

2. The Likelihood Gamble Pricing method

The Likelihood Gamble Pricing (LGP) Approach provides a new way to solve classification problems. In particular, we consider the problem of predicting the probability of the binary label (presence of side-effect, survival at 2-years) of a new example in the light of the available training data.

From a theoretical perspective, the LGP approach is a marriage of the likelihood principle of statistics with Von Neumann-Morgenstern's axiomatic approach to decision making, similar to Wald's seminal work in the 1940-50s (Wald, 1950).

In the LPG approach, the (labeled) training data and the (unlabeled) new example are combined in a principled manner as to achieve a regularized prediction of the probability of the unknown label. Interestingly, this regularized prediction is achieved by solving two (unregularized) maximum likelihood problems, see steps 1 and 3 in the procedure below.

2.1. Summary of Algorithm

This procedure is derived as an approximation to the likelihood pricing gamble approach developed in (Giang, 2006; Giang & Shenoy, 2005; Giang & Shenoy, 2002). The derivation is omitted due to lack of space in this application paper.

Consider training data $D = \{(y_i, x_i, w_i)\}_{i=1,\ldots,N}$ with N examples, where x_i are the input vectors, $y_i \in \{0, 1\}$ is the label, and $w_i \in \mathbb{R}^+$ is the weight of example i for training the classifier; $w_i = 1$ for all x_i in the set D. Given a new example $x_\diamondsuit$, the objective is to predict the probability $y_\diamondsuit^{\mathrm{f}}$ that its label is 1. While this method can be used for all generalized linear models with parameter vector β, we provide as a specific example the equations that apply to the logistic regression model. The (standard) log likelihood of the logistic model reads

$$\ell(\beta|D) = \sum_{i=1}^{N} w_i \left\{(y_i - 1)x_i\beta - \log(1 + \exp(-x_i\beta))\right\}. \tag{1}$$

The steps of our procedure are as follows:

1. determine $\hat{\beta}$ by solving standard maximum likelihood problem:

$$\hat{\beta} \quad = \quad \arg\max_{\beta} \ell(\beta|D) \qquad (2)$$

2. calculate weight $w_\diamondsuit$ assigned to the example $x_\diamondsuit$:

$$w_\diamondsuit = \frac{-\mathrm{sign}(x_\diamondsuit \hat{\beta})}{\frac{1}{1+\exp(x_\diamondsuit \hat{\beta})} - \frac{1}{2}} \qquad (3)$$

3. determine β^* by solving the maximum likelihood problem, where the new example $x_\diamondsuit$ is combined with the training data D using $y_\diamondsuit = 0.5$ and $w_\diamondsuit$ from the previous step:

$$\beta^* \quad = \quad \arg\max_{\beta} \ell(\beta|D \cup \{(y_\diamondsuit, x_\diamondsuit, w_\diamondsuit)\}) \quad (4)$$

4. now that $\hat{\beta}$ and β^* are determined, the 'fair' log likelihood ratio λ^{f} is calculated:

$$\lambda^{\mathrm{f}} = x_\diamondsuit \beta^* + \mathrm{sign}(x_\diamondsuit \hat{\beta}) \left[\ell(\hat{\beta}|D) - \ell(\beta^*|D) \right]. \quad (5)$$

5. finally, λ^{f} can be mapped to the 'fair' price $y^{\mathrm{f}}_\diamondsuit$, which corresponds to the predicted probability if the lable being 1:

$$y^{\mathrm{f}}_\diamondsuit = \frac{1}{1 + \mathrm{e}^{-\lambda^{\mathrm{f}}}} \qquad (6)$$

Note that step 1 needs to be computed only once, while steps 2 through 5 have to be computed for every new example $x_\diamondsuit$ to be classified. Note that $\hat{\beta}$ can serve as an excellent initialization for optimizing β^*.

If β^* is not close to $\hat{\beta}$, then the accuracy of our linear approximation can be improved by iterating steps 2 and 3 until convergence. In this case, use in step 2 the value β^* from the previous iteration in place of $\hat{\beta}$.

3. Experiments

3.1. Datasets and Problem Description

We apply our method on four real-world medical problems in this section. They are all challenging problems for radiation oncology.

3.1.1. NSCLC Survival Analysis

Our first problem is 2-year survival prediction for advanced non-small-cell lung cancer (NSCLC) patients treated with (chemo-)radiotherapy. This is currently a very challenging problem in clinical research, since the prognosis of this group of patients is very poor (less than 40% survive two years). At present, generally accepted prognostic factors for inoperable patients are performance status, weight loss, presence of comorbidity, use of chemotherapy in addition to radiotherapy, radiation dose and tumor size. In a recent study it was shown that number of involved nodal areas quantified by PET-CT was also an important prognostic factor for survival (Dehing-Oberije et al., 2007).

A total number of 460 inoperable NSCLC patients, stage I-IIIB, were referred to MAASTRO clinic to be treated with curative intent between May 2002 and January 2007. Treatment protocols prescribed the treatment regimen as well as the collection of relevant data, ensuring standardization of data collection and high quality of the data. For additional information clinical charts of the patients were reviewed. If PET was not used as a staging tool, patients were excluded from the study. To test our method in a classification setting, we label each patient as +1 if the survival time is at least 2 years, and −1 if it is not. All the right-censored patients are excluded in this study. This resulted in the inclusion of 322 patients. We consider the following 8 factors in the model: the gender of the patient, presence of comorbidity, the WHO performance status, forced expiratory volume in 1 second (FEV1) in the lung function test, number of positive lymph node stations, overall treatment time (in days), the equivalent total dose in the treatment, and the tumor size.

3.1.2. Radiation-Induced Lung Toxicity

In a second study we try to predict the radiation-induced lung toxicity (RILD) in radiotherapy for this set of NSCLC patients. It is generally accepted that risk of RILD depends on radiation dose as well as irradiated volume. Extensive research has lead to the identification of numerous dosimetric parameters associated with lung toxicity, but their clinical usefulness remains largely unknown. We decided to investigate the predictive value of the *mean lung dose* in combination with other treatment-related factors and patient characteristics

The same set of patients in the NSCLC survival study were used for this study. Lung toxicity was scored using the Common Toxicity Criteria version 3.0. Acute dyspnea was defined as dyspnea grade II or higher (i.e., as labeled +1 in classification), during or at maximum 6 months after radiotherapy treatment. The factors we considered in the model building consist of age, gender, presence of comorbidity, the WHO performance status, smoking history, FEV1, use of chemotherapy or not, the location of the tumor, and the mean lung dose.

3.1.3. Radiation-Induced Dysphagia

Another important side effect of (chemo-)radiotherapy for NSCLC patients is the dysphagia which is the radiation-induced damage to the esophagus. Many research groups have identified radiobiological factors such as the irradiated surface, the volume and length of the irradiated esophagus, the use of concurrent chemotherapy and the overall treatment time of radiotherapy. However, combined chemotherapy and radiotherapy risk factors for dysphagia in concurrent chemoradiation schedules are less well studied and are challenging. For this study we consider 18 various factors including gender, age, type of chemotherapy, overall treatment time, maximal dose to esophagus, and neutropenia.

3.1.4. Pathologic Complete Response (pCR) for Rectal Cancer

Our last example is to predict tumor response after chemo-radiotherapy for locally advanced rectal cancer. This is important in individualizing treatment strategies, since patients with a pathologic complete response (pCR) after therapy, i.e., with no evidence of viable tumor on pathologic analysis, would need less invasive surgery or another radiotherapy strategy instead of resection. Most available models combine clinical factors such as gender and age, and pre-treatment imaging-based factors such as tumor length and $\text{SUV}_{\max}$ (from CT/PET imaging), but it is expected that adding imaging data collected after therapy would lead to a better predictive model (though with a higher cost). In this study we show how our method performs with pre-treatment and post-treatment imaging features to predict pCR.

We use the data from (Capirci et al., 2007) which contains 78 prospectively collected rectal cancer patients. All patients underwent a CT/PET scan before treatment and 42 days after treatment, and 21 of them had pCR (labeled +1). The relevant factors we considered in the study include gender, age, maximal diameter of the tumor, staging of cancer, the $\text{SUV}_{\max}$ before treatment, and the absolute difference of $\text{SUV}_{\max}$ before and after treatment (in values and percentages).

3.2. Description of the experiments

Next, we present numerical comparison of the following three methods: standard (maximum-likelihood) Logistic Regression (**LR**), regularized Logistic Regression using a 2-norm penalty term (**RLR**) (Mukherjee et al., 2002), and our proposed likelihood gamble pricing approach (**LGP**) applied to the logistic regression model, see the procedure in Section 2.1. In this application paper, we present experiments on various medical data sets as described above. For each data set, we randomly partitioned the available data into two disjoint subsets, one of them was used for training (training set), the remaining data was used for testing. To investigate our suspicions that our algorithm would perform better than the standard formulation when the training data is scarce, we repeated our experiments over a wide range of sizes of training data (as a percentage of the entire available data set). For each of these random splits, the experiments were performed 10 times. For the **RLR** method the regularization parameter was chosen in the set $\left\{2^{-10}, 2^{-6}, \ldots, 2^{4}, 2^{5}\right\}$ by a cross-validation tuning procedure on the available training data.

We are interested in the *quantitative* prediction concerning the probability of the label being 1, rather than in the *qualitative* prediction whether the predicted label is 0 or 1 (which could be measured in terms of 0/1-loss). For this reason, we used the Kullback-Leibler divergence between the empirical distribution of the true labels and the predicted probabilities as a measure of performance. The Kullback-Leibler divergence provides a standard way of comparing two probability distributions, and is hence conceptually a fair measure for evaluating methods that aim at optimizing a (regularized) likelihood on the trainig data.[1] The values reported correspond to the mean and the standard deviation over five runs.

3.3. Experiments in medical data sets

As shown in Figure 1, the **LPG** approach consistently makes more robust predictions (less variance) that are closer to the true labels according to the distance induced by the empirical KL divergence. One important thing to note is that the regularization induced by the **LPG** approach produced better probabilistic predictions than both the unregularized and the regularized (by the standard 2-norm regularization) logistic regression approaches when the training dataset was small.

[1] Other measures, e.g., like the Area under the ROC Curve which assesses the quality of the predicted ranking, are conceptually different, and hence would not provide a fair evaluation of a likelihood-based method.

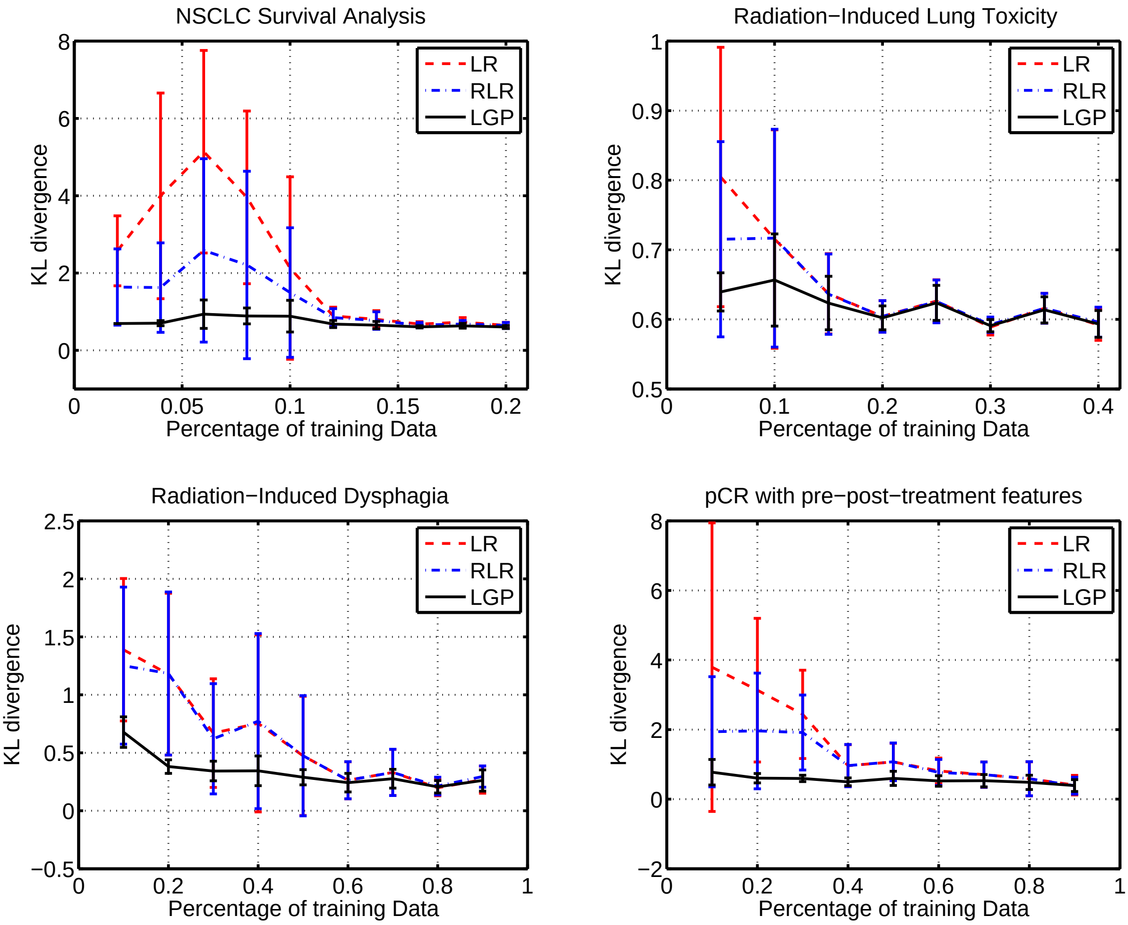

Figure 1. Comparison of three methods on four medical data sets: Logistic regression (**LR**), Regularized logistic regression (**RLR**) and the proposed likelihood gamble pricing approach (**LGP**). The mean and standard deviation of the KL divergence between the real and the predicted values are reported.

4. Conclusions

We have applied a recently proposed a general procedure for statistical forecasting to the problem of probability prediction for personalized medicine predictive models. The logistic gamble pricing (LGP) approach is fundamentally different from two traditional approaches namely model selection and Bayesian model averaging. In model selection, one model is selected as a proxy for the ideally true model and a forecast value is obtained by plug into the selected model the value of the predictor variable. Maximum likelihood and likelihood-based variations (AIC, BIC etc) are possible criteria for model selection. Model selection approach ignores predictions made by other plausible models. Bayesian model averaging approach averages the forecasts by individual models with posterior model probabilities that are computed from likelihood and a prior model probability. LGP is an attractive alternative to these traditional approaches. The main advantage of the LGP approach is that it can make use of model uncertainty information provided by data without prior probability assumption which is generally the case in medical settings where the data is scarce.

Experiments in several real-life medical data for survival and side-effects prediction for lung cancer and rectal cancer radiotherapy are performed. We compare the performance of LGP against both the standard logistic regression and the regularized logistic regression model. A potential cost to pay for the improved accuracy is the increase in computational cost because unlike model selection, LGP does not rely on any specific “learned” model to make predictions. In the original naive computational settings, each prediction by LGP involves solving two optimization problems that are comparable to learning a model. While the tradeoff between computation cost and accuracy is naturally expected (no free lunch principle), in most medical applications, the concern about forecast accuracy would overwhelm the computation cost. However in this paper, we use an approximation method that is fast for medium sized data sets. The numerical results show the superiority of the new method in the real-life data used, specially when the training data set is small.

Acknowledgements

The authors want to thank the following persons for their invaluable support in the making of this paper: Cary Dehing-Oberije, Krishnan Sriram, R. Bharat Rao, Philippe Lambin.

References

Capirci, C., Rampin, L., Erba, P., Galeotti, F., Crepaldi, G., Banti, E., Gava, M., Fanti, S., Mariani, G., Muzzio, P., & Rubello, D. (2007). Sequential FDG-PET/CT reliably predicts response of locally advanced rectal cancer to neo-adjuvant chemoradiation therapy. *Eur J Nucl Med Mol Imaging*, *34*.

Cox, D. R., & Oakes, D. (1984). *Analysis of survival data*. Chapman and Hall.

Dehing-Oberije, C., Ruysscher, D. D., van der Weide, H., & et al (2007). Tumor volume combined with number of positive lymph node stations is a more important prognostic factor than tnm stage for survival of non-small-cell lung cancer patients treated with (chemo)radiotherapy. *Int J Radiat Oncol Biol Phys*.

Drish, J. (2001). Obtaining calibrated probability estimates from support vector machines. *Technical report, http://citeseer.nj.nec.com/drish01obtaining.html.*

Giang, P. H. (2006). A new axiomatization for likelihood gambles. *Uncertainty in Artificial Intelligence: Proceedings of the 22nd Conference (UAI–2006)* (pp. 192–199). Corvallis, OR: AUAI Press.

Giang, P. H., & Shenoy, P. P. (2002). Statistical decisions using likelihood information without prior probabilities. *Uncertainty in Artificial Intelligence: Proceedings of the Eighteenth Conference (UAI–2002)* (pp. 170–178). San Francisco, CA: Morgan Kaufmann.

Giang, P. H., & Shenoy, P. P. (2005). Decision making on the sole basis of statistical likelihood. *Artificial Intelligence*, *165*, 137–163.

Mukherjee, S., Rifkin, R., & Poggio, T. (2002). Regression and classification with regularization. *Lectures Notes in Statistics: Nonlinear Estimation and Classification* (pp. 107–124).

Platt, J. (2000). Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. *Advances in Large Margin Classifiers* (pp. 61–74).

Szolovits, P., & Pauker, S. G. (1990). Categorical and probabilistic reasoning in medical diagnosis. 282–297.

Wald, A. (1950). *Statistical decision function*. New York: John Wiley and Sons.

Zadrozny, B., & Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive bayesian classiers. *Proceedings of the Eighteenth International Conference on Machine Learning, 2001.*

Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 694–699). New York, NY, USA: ACM.