Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence

Evaluating Trauma Patients: Addressing Missing Covariates with Joint Optimization

Alex Van Esbroeck[1], **Satinder Singh**[1], **Ilan Rubinfeld**[2], **Zeeshan Syed**[1]
[1]Computer Science & Engineering, University of Michigan, Ann Arbor, MI
[2]Henry Ford Hospital, Detroit, MI

Abstract

Missing values are a common problem when applying classification algorithms to real-world medical data. This is especially true for trauma patients, where the emergent nature of the cases makes it difficult to collect all of the relevant data for each patient. Standard methods for handling missingness first learn a model to estimate missing data values, and subsequently train and evaluate a classifier using data imputed with this model. Recently, several proposed methods have demonstrated the benefits of jointly estimating the imputation model and classifier parameters. However, these methods make assumptions that limit their utility with many real-world medical datasets. For example, the assumption that data elements are missing at random is often invalid. We address this situation by exploring a novel approach for jointly learning the imputation model and classifier. Unlike previous algorithms, our approach makes no assumptions about the missingness of the data, can be used with arbitrary probabilistic data models and classification loss functions, and can be used when both the training and testing data have missing values. We investigate the utility of this approach on the prediction of several patient outcomes in a large national registry of trauma patients, and find that it significantly outperforms standard sequential methods.

Introduction

Missing values occur frequently in medical datasets for a variety of reasons: sensors may malfunction, equipment may be unavailable, or a patient may not be in a condition where certain measurements can be recorded. While this is an issue for many medical applications, it is particularly troublesome when stratifying trauma patients for adverse outcomes. Trauma is the leading cause of death under the age of 44 years in the United States, and costs approximately $80 billion in medical treatment a year (Finkelstein, Corso, and Miller 2006). Trauma patients require urgent care and are often in very poor condition at their arrival to the hospital. As a result, it is generally impossible to collect all of the potentially relevant clinical measurements at the time of admission, and this missingness cannot be reduced through more staff or equipment as is the case in other domains. This prevalence of missing values is a significant obstacle to the application of machine learning in trauma care, as the vast majority of methods do not work with missing values.

There is an extensive literature within the statistics community on robustly estimating statistics from incomplete data. The most commonly used methods rely on inferring one or more models of the observed data distribution, e.g., by using expectation maximization (EM). Once the parameters of the distribution have been learned, these models can be used to impute the missing values, or reason about their uncertainty when estimating statistics from the data.

Recently, studies in the machine learning community have investigated the handling of missing values in the context of classification. These methods can be organized into several different categories. Some methods eschew imputation models entirely, attempting to learn classifiers that can robustly handle inputs with missing values without attempting to fill them in (Chechik et al. 2007; Grangier and Melvin 2010). Others learn an imputation model using standard methods (such as EM), and consider the full distribution over possible values of missing attributes when training the classifier (Williams et al. 2007; Smola, Vishwanathan, and Hoffman 2005). A third direction, and the focus of the present work, jointly learns the parameters of the imputation model and the classifier (Liao, Li, and Carin 2007; Dick, Haider, and Scheffer 2008; Wang et al. 2010).

This joint learning of the imputation model and classifier has been shown to be useful for a variety of reasons (Dick, Haider, and Scheffer 2008). First, sequential optimization (learning the imputation model first, followed by the classifier) is prone to compounding errors: mistakes made in learning the imputation model parameters are propagated into the training of the classifier. Joint optimization can avoid this issue by taking into account the effect of the imputation model on classification performance when estimating the imputation model parameters. Second, different choices of imputation models may be better suited to different classifiers, as some classifiers may be better at handling certain kinds of errors than others.

While several methods have been developed to jointly learn the imputation model and classifier parameters, they have some critical limitations that prevent applicability to real-world datasets. Existing methods either do not work when both training and test data have missing values (Dick,

Haider, and Scheffer 2008), or assume that the data is missing at random (Liao, Li, and Carin 2007), which is unlikely to hold true in many real-world medical settings.

In this paper, we addresses these limitations and extend the applicability of joint optimization of the imputation model and classifier to a broad set of medical datasets. Our proposed method is an optimization problem over both imputation model and classifier parameters, which uses the effect of the imputation model on classification loss to guide the solution towards imputation parameters that achieve better classification performance. Unlike existing methods for joint optimization, the proposed method makes no assumptions about the missingness mechanism of the data. This makes it better suited to data that is not missing at random (NMAR), and allows it to be used when both training/testing data have missing values. Our method can be used with a variety of choices of imputation model or classification loss functions and allows the joint optimization of probabilistic imputation models with discriminative classifiers.

The contributions of this paper are as follows:

- We present a novel optimization problem for evaluating trauma patients that simultaneously considers imputation and classification. Unlike existing literature, our approach is broadly applicable to medical datasets.
- We study the utility of jointly learning imputation and classification parameters on synthetic data and quantify how such a method provides benefits for NMAR data. NMAR data has received little attention in prior work but may be more common in clinical applications.
- We compare the proposed method for joint optimization to standard sequential learning when predicting several important patient outcomes in a large national registry of trauma patients. We demonstrate in a representative cohort of patients that our method provides significant improvement across several metrics.

Background

The canonical definition of missingness mechanisms comes from Little and Rubin (1987). Consider a dataset X, with all values present. The missingness matrix R is defined to be the same size as X, where

$$R_{ij} = \begin{cases} 0 & \text{if } X_{ij} \text{ is missing} \\ 1 & \text{if } X_{ij} \text{ is observed} \end{cases} \quad (1)$$

Define X^o as the portion of X that is observed (where $R = 1$), and X^m as the portion of X that is unobserved. The joint distribution of X and R can be modeled as

$$P(X, R|\theta, \phi) = P(X|\theta)P(R|X, \phi) \quad (2)$$

where θ are parameters governing the data distribution, and ϕ are parameters governing the missingness R, which may depend on the values in the data X. The missing completely at random (MCAR) mechanism assumes that the missingness is independent of the data:

$$P(R|X, \phi) = P(R|\phi) \quad (3)$$

A weaker assumption, and one that subsumes MCAR, is the missing at random (MAR) mechanism, which allows the missingness to depend on the observed, but not the unobserved values of X

$$P(R|X, \phi) = P(R|X^o, \phi) \quad (4)$$

When even this this MAR condition does not hold, the data is said to be not missing at random (NMAR). In the NMAR setting, the distribution of a missing variable conditioned on the observed variables may differ from the conditional distribution of that variable had it been observed. This makes it impossible to estimate the distribution from the data without information on the missingness mechanism. For example, if older respondents to a patient survey are less likely to report their age, the average of the missing age values will be higher than that of the observed age values. In the NMAR setting it is difficult to learn good estimates of the imputation model parameters θ, as the distributions of the data X and missingness R become coupled, requiring an explicit model of the missingness process.

Imputing Missing Values

Many methods exist for imputing missing values (García-Laencina, Sancho-Gómez, and Figueiras-Vidal 2010). These range from simple approaches like filling missing values in with the mean of the observed data, to more complex approaches based on weighted averaging of the k-nearest neighbors or random forests.

The most commonly used approaches for imputation rely on statistical models of the data and use data distribution parameters to either fill in missing values with point estimates (e.g., their expected value conditioned on the observed values), or to generate many samples that approximate the distribution over possible values (multiple imputation). The EM algorithm is the most frequently used method for learning these data distribution parameters in the presence of missing values. Given a flexible class of models, such as the commonly used Gaussian mixture model, EM can easily learn models of the data distribution (Williams et al. 2005; Ghahramani and Jordan 1994). Most applications of EM to imputation take advantage of the MAR assumption, seeking to optimize the full data likelihood over choices of model parameters by assuming the missing and observed data share the same distribution. It is clear in the case of NMAR that the model parameters that maximize the observed data likelihood may differ, possibly substantially, from the true parameters. While it is possible to use EM without making the MAR assumption, it requires defining a model for the generation of missingness in the data, which is challenging as the mechanism is often prohibitively complex or unknown.

One straightforward approach to handling NMAR data is the inclusion of missingness indicator variables: constructing R from the data and adding this missingness matrix to the set of predictive variables. This can improve performance when missingness is related to the class labels, however it does not help find a better value for θ, and the addition of missingness indicators cannot in general correct the noise induced by an erroneous imputation model.

In this work we focus on single imputations, however multiple imputation methods are also used in practice (Sterne et al. 2009; Buuren and Groothuis-Oudshoorn

2011). However, as with EM-based single imputations, avoiding the MAR assumption with these methods usually requires making explicit assumptions about the missingness mechanism behind the data, and they ignore the choice of classifier or classification performance when learning the model parameters.

Joint Learning of Imputation Model and Classifier

In the standard sequential approach to learning imputation model and classifier, the imputation parameters θ are first learned using EM, and a classifier is then trained given the learned values of θ. This process may involve using point estimates of the missing values (their expected values under θ conditioned on the observed values), or using a classifier designed to account for the conditional distribution over missing values (Williams et al. 2007; Smola, Vishwanathan, and Hoffman 2005).

In contrast, a joint learning of the imputation model and classifier can help find better choices of θ. Inaccuracies in imputation add noise to the classification task reducing accuracy. By accounting for classification loss when learning the imputation model during joint optimization, we can therefore avoid the pitfalls of a sequential approach (e.g., compounding errors, ignorance of classifier choice).

Liao et al. developed a graphical model which incorporated both the data distribution and the classification task (2007). When estimating this unified model's parameters using EM, the imputation model and classifier are learned together. This method uses a monolithic probabilistic model, and cannot be applied with different classification loss functions, like hinge loss or different regularization terms. Because this unified probabilistic model does not explicitly model the missingness mechanism in the data, it depends upon the MAR assumption, limiting the method's applicability to NMAR data.

Dick et al. proposed a method that learns an exact imputation of the missing values in the training data using an objective that takes the classification loss into account (2008). This method allows for flexibility in choice of classification loss function, and does not make the MAR assumption. Unfortunately, it does not learn a parametric imputation model, but rather an exact imputation for the training data. As a result, the method is unusable when there are missing values in the testing data. Complete test data is frequently unavailable in many medical scenarios, precluding use of the method in such cases.

We note that when the MAR condition is violated, maximizing the observed data likelihood may not help (and could even hurt) the accuracy of the imputed values. In the context of classification however, we have an additional metric (loss function) that can aid in learning good imputation models, and improve our handling of NMAR data by relying less on the observed data likelihood. A more accurate imputation model introduces less noise to the classification process, and allows for better predictions. For this reason, the effect of the imputation model on classification loss can help guide our choice of imputation parameters, even when we have no information about the missing data distribution.

Methods

We propose a joint optimization problem that does not make the MAR assumption, can be used when both training and testing data have missing values, and can be used with a variety of imputation models and classification loss functions. The proposed joint optimization problem is presented in Equation 5.

$$\underset{\theta,w}{\operatorname{argmax}}(1-\alpha)LL(X^o|\theta) - \alpha Loss(Y|X^o,\theta,w) \quad (5)$$

Equation 5 seeks to maximize a convex combination of the observed data log likelihood under the imputation model parameters θ (the first term), and the classification loss using both the classifier parameters w and the data imputation under θ conditioned on the observed data X^o (the second term), where Y denotes the class labels for the data in X. Unlike with a unified probabilistic model, learning the parameters in this combined objective function does not require making probabilistic assumptions about the missingness mechanism of the data, as in the method of Liao et al. (2007), because the second term in the optimization is not probabilistic. However, in contrast to the method of Dick et al. (2008) we assume the existence of a parametric imputation model, described by θ, which allows us to use the learned model parameters to generate imputations on unseen data with missing values.

The tuning parameter $\alpha \in [0,1]$ in Equation 5 controls the relative strength of the classification loss and observed data likelihood in learning the parameters. In the case where α is (nearly) equal to zero, the method is equivalent to the traditional sequential optimization approach. The values of θ are determined entirely by the observed data likelihood, as in EM, and the classifier parameters w are optimized given that value of θ. When $\alpha = 1$, the classification loss alone guides the choice of parameters, effectively assuming that the observed data likelihood has no relevance. In this case, Equation 5 will try to find an imputation model that maximizes the separability of the data.

Decreasing the value of α can be thought of as a kind of regularization. Total reliance on classification loss to learn both the classifier and imputation model, which may have many parameters (particularly in high-dimensional data), is likely to lead to overfitting in smaller datasets. Reducing α encourages the imputation model to be closer to the observed data distribution. This can be loosely interpreted as treating the EM solution as a kind of prior on the missing data distribution, with α controlling the strength of that prior belief.

Equation 5 can be used with a variety of loss functions (e.g. log loss, hinge loss, or squared loss in the case of regression), as well as different data models (e.g. GMMs, multinomial mixture models). This allows for the combination of probabilistic imputation models with discriminative classifiers, and allows flexibility in the kind of regularization used for the classifier.

We use alternating optimization of θ and w to optimize Equation 5, as shown in Algorithm 1. After generating intial estimates of the imputation model θ_0 and the classifier

Algorithm 1 Alternating optimization of Equation 5

Input: X, Y, α, K
randomly initialize θ_0 (with K components)
w_0 = TrainClassifier(X^o, $E[X^m|X^o, \theta_0]$, Y)
$i = 0$
repeat
  $i = i + 1$
  $\theta_i = \text{argmax}_\theta (1-\alpha) LL(X^o|\theta) - \alpha Loss(Y|X^o, \theta, w_{i-1})$
  w_i = TrainClassifier(X^o, $E[X^m|X^o, \theta_i]$, Y)
until θ and w have converged
Output: θ_i, w_i

w_0, we alternate between optimizing the imputation model parameters θ conditioned on the current estimate of w, and optimizing the classifier parameters w using the imputation generated by the current imputation parameter estimate θ. The process repeats until the parameters have converged.

Algorithm 1 treats the second term of Equation 5 as the loss given a single imputation ($E[X^m|X^o, \theta]$). Using the expected value of the missing data under the imputation model parameters θ makes optimization of the second term comparable to training the classifier in the normal setting. However, it is possible to treat this term as the loss with integration over possible values of X^m, as investigated in several works (Williams et al. 2007; Smola, Vishwanathan, and Hoffman 2005).

We implemented Algorithm 1 using l_2 regularized log loss (logistic regression) for the classifier, and use GMMs for the data model. We optimized Equation 5 with respect to the GMM parameters θ given the classification weight vector w using gradient ascent. Estimation of w given θ was done using standard methods for training a logistic regression model, after filling in the missing values in X with their expected values. A validation set was used to select appropriate choices of α, the number of GMM components K, and the regularization parameter for the classifier.

The optimization of Equation 5 is susceptible to local optima. As a result, we run the optimization with multiple random parameter initializations, and select the best model/classifier on a validation set.

We evaluated our proposed method on both synthetic and real-world data. In what follows, we first present results on a synthetic dataset comparing sequential optimization using EM-based imputations with the proposal joint optimization algorithm on artificially generated missingness varying from MCAR to NMAR. This serves as an illustrative example of the method, and as an investigation into the relationship between α, the missingness mechanism, and the method's performance.

For our primary evaluation of the method, we compare sequential and joint optimization in their abilities to predict several adverse patient outcomes in a large national representative cohort of trauma patients. This evaluates the merits of the approach on real-world hospital data in several important prediction tasks.

Synthetic Data Evaluation

Synthetic data were used to compare the effect of missingness mechanism (MCAR vs. NMAR) on the method's performance, and particularly on the optimal choice of α. 1,000 data points were sampled from a 2-dimensional, 2-component mixture of Gaussians. The labels were generated by a perfect linear separator (see left panel in Figure 1).

Two different missingness mechanisms were used to generate the data. For the MCAR case, values were removed from the variable x_2 from all points with equal probability. To generate NMAR data, values were removed from the variable x_2 only in points generated by one of the components. Points generated by the other component had no missing values. As a result, the joint distribution of the variables x_1 and x_2 differed between the fully and partially observed data points.

To assess the effect of the amount of missing data on the method's performance, we varied the percentage of data points in which x_2 was missing. Once this percentage reaches 50% in the NMAR case, x_2 is entirely unobserved for points generated by the component with missing values, and the choice of imputation model becomes irrelevant. As a result, we limited the maximum percentage of points missing values for x_2 to 40%.

We compared the performance over a range of choices of α between 0 (equivalent to sequential optimization using EM) and 1 (fully loss-based approach), to see whether an objective based on classification loss could select a good imputation model and classifier. The two methods were compared over 20 random splits of the data into training (60%), validation (20%), and testing (20%) sets. Classification performance was measured using the area under the receiver operating characteristic curve (AUC), with the reported statistics reflecting the average across trials.

Results

Figure 1 shows the data used in the experiments (left panel), as well as the effect of the choice of α on AUC for both NMAR and MCAR missingness mechanisms (center and right panels). When the missingness was generated MCAR, there was a very minor improvement of α values greater than zero over the baseline sequential case ($\alpha = 0$). This is consistent with earlier work on joint optimization, which found small but statistically significant improvements in classification on MCAR data (Dick, Haider, and Scheffer 2008).

In the NMAR setting, α values greater than zero showed much larger improvements than in the MCAR setting. The improvement of joint optimization ($\alpha > 0$) increased with the amount of missing data, becoming particularly pronounced with 30% or more values missing. The loss term of Equation 5 is more sensitive to overfitting when fewer points have missing values, as it is determined by only a small number of data points. This justifies the use of smaller values of α with low amounts of missing data.

Prediction of Trauma Patient Outcomes

We investigated the performance of our method when predicting several adverse outcomes in a representative co-

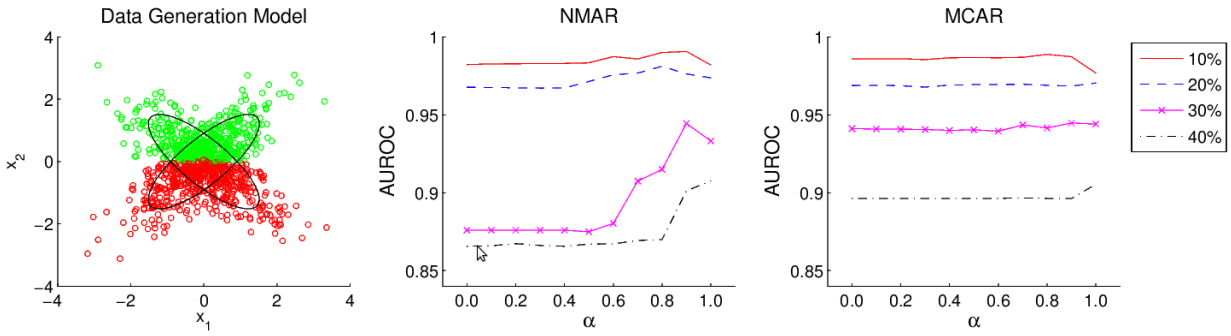

Figure 1: (Left) Depiction of the data used in the synthetic experiments, along with the generating GMM. The color of data points denotes class labels. (Center) Classification performance using different choices of the tuning parameter α on data with NMAR missingness, shown for varying percentages of missingness. (Right) Same as center, using MCAR missingness.

hort of trauma patients. The National Trauma Data Bank (NTDB) collects information about patients and outcomes from trauma centers around the country. The NTDB National Sample Program data from 2009 was used under IRB approval and the data use agreement of the American College of Surgeons. The dataset consisted of 162,821 records. The variables used in the analyses included a variety of vital signs and scores collected at admission to the emergency department: systolic blood pressure, pulse rate, respiration rate, oxygen saturation, temperature, injury severity score, and the Glasgow coma scale. The percentage of values missing for each attribute ranged from 3% to 40%. The missingness of variables in NTDB is typically due to an inability to collect these variables due to the patient's condition and therefore meets the requirement of NMAR.

The outcomes of interest were whether a patient was admitted to the ICU (34% of patients), whether they were put on a ventilator (17% of patients), and whether they suffered in-hospital mortality (3% of patients). For evaluation, the dataset was randomly divided into equal sized training, validation, and testing sets. The validation set was used to select α, the number of GMM components, and the classification regularization parameter. For each clinical feature an indicator variable (0 or 1) was added to the training data to account for whether this variable was available or missing. Evaluation was conducted both with and without use of these missingness indicator variables to assess whether the proposed method provides complementary improvements for NMAR data beyond simply knowing that there was no opportunity to collect certain measurements.

The goal of learning models in this setting is to predict outcomes on new trauma patients coming into the hospital with missing values. The approach of Dick et al. cannot be used when test data has missing values, and would not be useable on more than 50% of the patients in the test set. As a result, we compare the proposed approach against standard sequential optimization using the EM algorithm, which represents the most meaningful baseline in this setting.

Classification performance was measured using the area under the receiver operating characteristic curve (AUC), and the significance of an improvement in AUC was measured using the method of Delong et al. (1988). Additionally, the categoryless net reclassification improvement (NRI) and integrated discrimination improvement (IDI) were used to assess the improvement of using joint optimization instead of sequential optimization (Pencina et al. 2008). Use of these metrics has become widespread in the medical literature when assessing the improvement of one risk model over another.

NRI measures the proportion of patients x whose estimated risk probabilities under a new model, $\hat{p}_{new}(x)$, are more accurate than estimates from an old model, $\hat{p}_{old}(x)$. The direction of change in estimated risk, $v(x_i)$, for a patient x_i is defined as follows:

$$v(x_i) = sign(\hat{p}_{new}(x_i) - \hat{p}_{old}(x_i)) \tag{6}$$

The NRI combines the percentage of patients with events $[y_i = 1]$ whose risk scores increase under the new model, with the percentage of patients without events $[y_i = 0]$ whose risk scores decrease:

$$NRI = \frac{\sum_{i,\, y_i=1} v(x_i)}{\sum_k [y_k = 1]} - \frac{\sum_{j,\, y_j=0} v(x_j)}{\sum_k [y_k = 0]} \tag{7}$$

IDI instead focuses on the magnitude of improvements in the estimated risk probabilities:

$$IDI = \frac{\sum_{i,\, y_i=1} \hat{p}_{new}(x_i) - \hat{p}_{old}(x_i)}{\sum_k [y_k = 1]} - \frac{\sum_{i,\, y_i=0} \hat{p}_{new}(x_i) - \hat{p}_{old}(x_i)}{\sum_k [y_k = 0]} \tag{8}$$

Unlike AUC, which considers only the ordering of data points, these metrics account for the accuracy of the predicted probabilities, which is of great practical importance in the medical domain.

Results

Table 1 shows the performance of various approaches on several patient outcomes on the NTDB dataset. The proposed joint optimization method achieved statistically significant higher AUC values than sequential optimization with mean imputation and EM imputation in predicting ICU

	No Indicators			Indicators		
	Seq. (Mean)	Seq. (EM)	Joint	Seq. (Mean)	Seq. (EM)	Joint
ICU	0.676	0.681 (*ref.*)	**0.691 (*0.004*)**	0.684	0.685 (*ref.*)	**0.696 (*0.002*)**
Ventilator	0.757	0.783 (*ref.*)	**0.794 (*0.007*)**	0.794	0.793 (*ref.*)	**0.798 (*0.094*)**
Mortality	0.833	0.831 (*ref.*)	0.835 (*0.303*)	0.837	0.836 (*ref.*)	0.836 (*0.513*)

Table 1: AUC values for three outcomes on the NTDB dataset when using the sequential method with mean imputation [Seq. (Mean)], with EM imputation [Seq. (EM)], and the joint optimization method. P-values corresponding to the improvement of joint optimization over sequential are included in parentheses, with the baseline marked with (*ref.*). Results are shown with and without the use of missingness indicator variables. Bolded results indicate significant improvement at the 0.1 level.

	IDI (p)		NRI (p)	
	No Indicators	Indicators	No Indicators	Indicators
ICU	0.007 (*<0.001*)	0.008 (*<0.001*)	0.120 (*<0.001*)	0.245 (*<0.001*)
Ventilator	0.014 (*<0.001*)	0.005 (*<0.001*)	0.574 (*<0.001*)	0.397 (*<0.001*)
Mortality	0.001 (*<0.001*)	0.000 (*0.508*)	0.559 (*<0.001*)	0.268 (*<0.001*)

Table 2: Integrated discrimination improvement (IDI) and net reclassification improvement (NRI) scores and associated p values when assessing the change from sequential to joint optimization. Positive IDI and NRI values indicate an improvement in the accuracy of assigned probabilities by using joint instead of sequential optimization.

admission and ventilator use. This improvement was consistent even after adding missingness indicator variables to the model. No significant difference was found between any of the methods in predicting in-hospital mortality. The presence of an improvement when using missingness indicator variables confirms that the missingness is likely to be NMAR.

Table 2 shows the IDI and NRI of using the proposed method over the EM-based sequential method. The use of joint optimization led to statistically significant improvements in IDI for all three outcomes of interest when missingness indicators were not used. When including missingness indicators, joint optimization had a statistically significant IDI for prediction of ICU admission and ventilator use, although not for mortality. Joint optimization achieved statistically significant NRI values for all outcomes, regardless of whether missingness indicators were included in the models or not. These results indicate that the use of our method resulted in more accurate risk probability estimates than with the standard approach to handling missingness.

Computational Overhead Joint optimization converged in fewer than 5 iterations (about 15 minutes to train on 50,000 examples using a 4-core Intel Xeon processor), taking approximately 20 times as long as the EM-based sequential method. We emphasize that the joint optimization training can be done offline with online evaluation of new patients taking the same amount of time as with standard methods.

Discussion

In this paper we present a general optimization problem that can jointly address imputation and classification. In contrast to existing methods for joint optimization, our method has several properties that make it applicable to real-world datasets: the proposed approach does not assume the data is MAR, is applicable when both training and testing data have missing values, and can use a variety of imputation models and classification loss functions.

Our method is motivated by the problem of evaluating trauma patients, for whom clinical variables frequently go uncollected due to the severity and urgency of their condition. When evaluated in a large and representative population of patients undergoing trauma surgery, our proposed approach achieved statistically significant improvements over standard sequential optimization in terms of the AUC, IDI, and NRI statistics, even with the addition of missingness indicator variables. While small in magnitude for some of the outcomes, even minor improvements in individual risk estimates could have a substantial effect on patient care and quality and outcomes initiatives (potentially affecting tens of thousands of patients per year).

Our evaluation showed greater benefits when using joint optimization on NMAR data in comparison to the commonly studied MCAR data on datasets with artificially generated missingness. These results confirm the intuition that as the distribution of the missing data differs from that of the observed, the MAR assumption harms the classification performance, and that the inclusion of classification loss can better reflect the utility of a choice of imputation model parameters.

The method was evaluated using single imputation for estimation of missing values, however future work could extend it to work with multiple imputations. The approach could also be used for regression, for example, by using squared error as the loss function in the joint optimization.

Acknowledgments. This work was supported by NSF grant SHB 1064948. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the views of the sponsors.

References

Buuren, S., and Groothuis-Oudshoorn, K. 2011. Mice: Multivariate imputation by chained equations in r. *Journal of statistical software* 45(3).

Chechik, G.; Heitz, G.; Elidan, G.; Abbeel, P.; and Koller, D. 2007. Max-margin classification of incomplete data. *Advances in Neural Information Processing Systems* 19:233.

DeLong, E.; DeLong, D.; and Clarke-Pearson, D. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 837–845.

Dick, U.; Haider, P.; and Scheffer, T. 2008. Learning from incomplete data with infinite imputations. In *Proceedings of the 25th International Conference on Machine Learning*, 232–239. ACM.

Finkelstein, E. A.; Corso, P. S.; and Miller, T. R. 2006. *The incidence and economic burden of injuries in the United States.* Oxford University Press.

García-Laencina, P.; Sancho-Gómez, J.-L.; and Figueiras-Vidal, A. 2010. Pattern classification with missing data: a review. *Neural Computing and Applications* 19(2):263–282.

Ghahramani, Z., and Jordan, M. 1994. Supervised learning from incomplete data via an em approach. In *Advances in Neural Information Processing Systems*.

Grangier, D., and Melvin, I. 2010. Feature set embedding for incomplete data. In *Advances in Neural Information Processing Systems*.

Liao, X.; Li, H.; and Carin, L. 2007. Quadratically gated mixture of experts for incomplete data classification. In *Proceedings of the 24th International Conference on Machine Learning*, 553–560. ACM.

Little, R., and Rubin, D. 1987. *Statistical analysis with missing data*, volume 4. Wiley New York.

Pencina, M.; D'Agostino Sr, R.; D'Agostino Jr, R.; and Vasan, R. 2008. Evaluating the added predictive ability of a new marker: from area under the roc curve to reclassification and beyond. *Statistics in Medicine* 27(2):157–172.

Smola, A.; Vishwanathan, S.; and Hoffman, T. 2005. Kernel methods for missing variables. In *Advances in Neural Information Processing Systems*.

Sterne, J. A.; White, I. R.; Carlin, J. B.; Spratt, M.; Royston, P.; Kenward, M. G.; Wood, A. M.; and Carpenter, J. R. 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal* 338.

Wang, C.; Liao, X.; Carin, L.; and Dunson, D. 2010. Classification with incomplete data using dirichlet process priors. *The Journal of Machine Learning Research* 11:3269–3311.

Williams, D.; Liao, X.; Xue, Y.; and Carin, L. 2005. Incomplete-data classification using logistic regression. In *Proceedings of the 22nd International Conference on Machine Learning*, 972–979. ACM.

Williams, D.; Liao, X.; Xue, Y.; Carin, L.; and Krishnapuram, B. 2007. On classification with incomplete data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 29(3):427–436.