Artificial Intelligence in Medicine 72 (2016) 1–11

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Survival analysis for high-dimensional, heterogeneous medical data: Exploring feature extraction as an alternative to feature selection

Sebastian Pölsterl[a,*], Sailesh Conjeti[a], Nassir Navab[a,b], Amin Katouzian[c]

[a] Computer Aided Medical Procedures, Technische Universität München, Boltzmannstraße 3, 85748 Garching bei München, Germany
[b] Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218, USA
[c] IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120, USA

ARTICLE INFO

Article history:
Received 22 February 2016
Received in revised form 15 June 2016
Accepted 25 July 2016

Keywords:
Feature extraction
Feature selection
Dimensionality reduction
Survival analysis
Censoring
Spectral embedding

ABSTRACT

Background: In clinical research, the primary interest is often the time until occurrence of an adverse event, i.e., survival analysis. Its application to electronic health records is challenging for two main reasons: (1) patient records are comprised of high-dimensional feature vectors, and (2) feature vectors are a mix of categorical and real-valued features, which implies varying statistical properties among features. To learn from high-dimensional data, researchers can choose from a wide range of methods in the fields of feature selection and feature extraction. Whereas feature selection is well studied, little work focused on utilizing feature extraction techniques for survival analysis.
Results: We investigate how well feature extraction methods can deal with features having varying statistical properties. In particular, we consider multiview spectral embedding algorithms, which specifically have been developed for these situations. We propose to use random survival forests to accurately determine local neighborhood relations from right censored survival data. We evaluated 10 combinations of feature extraction methods and 6 survival models with and without intrinsic feature selection in the context of survival analysis on 3 clinical datasets. Our results demonstrate that for small sample sizes – less than 500 patients – models with built-in feature selection (Cox model with ℓ_1 penalty, random survival forest, and gradient boosted models) outperform feature extraction methods by a median margin of 6.3% in concordance index (inter-quartile range: [−1.2 %;14.6 %]).
Conclusions: If the number of samples is insufficient, feature extraction methods are unable to reliably identify the underlying manifold, which makes them of limited use in these situations. For large sample sizes – in our experiments, 2500 samples or more – feature extraction methods perform as well as feature selection methods.

1. Introduction

Medical data, such as electronic health records, often consist of a large set of heterogeneous variables, collected from different sources, such as demographics, disease history, medication, allergies, biomarkers, medical images, or genetic markers; each of which offers a different partial view on a patient's state. Moreover, statistical properties among aforementioned sources are inherently different: information about a patient's disease history is often obtained in form of a questionnaire, whereas biomarker measurements denote the concentration of metabolites in the blood; the first is categorical, whereas the second is continuous valued. When analyzing such data, researchers and practitioners are confronted with two problems: (1) the curse of dimensionality – the number of samples required to adequately sample the feature space is increasing exponentially in the number of dimensions – and (2) the heterogeneity in features' sources and statistical properties.

In this paper, we focus on two general groups of algorithms for dimensionality reduction, namely *feature selection* and *feature extraction*, and investigate how well these algorithms perform in a wide range of scenarios. We evaluated 10 combinations of feature extraction methods and 6 survival models with and without intrinsic feature selection in the context of survival analysis on 3 clinical datasets. The key finding of our experiments is that the combination of a linear survival model with spectral embedding methods can achieve performance on par with non-linear survival models with embedded feature selection, if the available

* Corresponding author.
E-mail addresses: sebastian.poelsterl@tum.de (S. Pölsterl), conjeti@in.tum.de (S. Conjeti), nassir.navab@tum.de (N. Navab), akatouz@us.ibm.com (A. Katouzian).

http://dx.doi.org/10.1016/j.artmed.2016.07.004
0933-3657/

number of samples is sufficient to reliably identify the underlying manifold. For insufficient sample sizes, feature extraction methods suffer from instabilities and survival models generalize poorly, in which case models with embedded feature selection are preferred. Next, we will briefly describe feature extraction, survival analysis, and review prior work on dimensionality reduction for survival analysis.

1.1. Feature extraction methods

Feature extraction methods construct a new, smaller set of features by (non-)linearly combining existing features. In contrast, feature selection methods assign each feature a value of importance, which is used to filter the set of features. Whereas feature selection methods have been well established in survival analysis, little work investigated the vast amount of feature extraction methods for survival analysis. Many feature extraction methods were originally proposed for computer vision problems [1–5], which often comprise more than 100,000 features and samples.

Most feature extraction methods are based on spectral decomposition and therefore all require the construction of a matrix that encodes global and/or local relations between data points. Principal component analysis (PCA) follows this scheme by computing an eigenvalue decomposition of a $p \times p$ covariance matrix, which encodes relationships between samples on a global scale. The resulting eigenvectors form the basis of a new space, whose dimensionality can be limited by only selecting $d < p$ eigenvectors corresponding to the d largest eigenvalues. Other techniques consider each sample's local neighborhood, which is encoded in a $n \times n$ neighborhood graph. A common choice to measure locality is a k-nearest neighbor search based on the Euclidean distance between samples. Using the neighborhood graph, the goal is to find a projection of the data to a low-dimensional space that preserves local neighborhoods as defined in the high-dimensional space (the process is also referred to as *low-dimensional embedding*). Laplacian eigenmaps (LE) [6] results in a non-linear transformation of the data, whereas locality preserving projections (LPP) [3] in a linear transformation. In both cases, a low-dimensional representation can be obtained by limiting the number of eigenvectors after spectral decomposition of the (normalized) graph Laplacian associated with the neighborhood graph. However, constructing the neighborhood graph based on the Euclidean distance is unsuitable when samples are medical records, because the Euclidean distance does not account for varying statistical properties among features. In addition, feature extraction methods mentioned above assume that feature vectors originate from a common vector space – they are called *singleview* spectral embedding methods. Thus, singleview algorithms are not aware of distinct sources of information and statistical properties they imply – which vary heavily in the case of medical records.

Dimensionality reduction in the presence of multiple independent groups of features with distinct statistical properties (called *views*) has been addressed by *multiview* spectral embedding (MVSE) [5]. Multiview spectral embedding first constructs a low-dimensional representation using Laplacian eigenmaps [6] for each view independently, followed by a global coordinate alignment to ensure that low-dimensional embeddings in different views are consistent with each other in the global context. The result is a non-linear transformation from the original high-dimensional feature space – comprising all views – to a low-dimensional space that preserves local neighborhoods of samples. A linearization of the objective function used in MVSE was proposed by Li et al. [7]. In addition, Liu et al. [8] took a sample's class label into account when constructing the neighborhood graph such that only samples of the same class are connected to each other.

1.2. Survival analysis

In many clinical studies, the primary interest is *survival analysis*, i.e., to establish a connection between a set of features and the time until an event of interest, such as death or emergence of a disease. Since patients can only be followed for a limited time and some patients choose to leave the study, survival times are *right censored*. Only if a patient experiences an event during the study period, one can record the exact time of the event, otherwise one only knows that the patient remained event-free during the study period and it is unknown whether an event has or has not occurred after the study ended. Consequently, survival data demands for algorithms that take this unique characteristic into account.

Given a dataset $\mathcal{D}$ of n patients, let $\mathbf{x}_i$ denote a d-dimensional feature vector, $t_i > 0$ the time of an event, and $c_i > 0$ the time of censoring of the ith patient. Due to right censoring, it is only possible to observe $y_i = \min(t_i, c_i)$ and $\delta_i = I(t_i \leq c_i)$ for every patient, with $I(\cdot)$ being the indicator function and $c_i = \infty$ for uncensored records. Hence, training a survival model is based on a set of triplets: $\mathcal{D} = \{(\mathbf{x}_i, y_i, \delta_i)\}_{i=1}^{n}$. After training, a survival model ought to predict a risk score of experiencing an event based on a set of features. Next, we will review work on dimensionality reduction in the biomedical domain.

1.3. Dimensionality reduction for biomedical applications

Dimensionality reduction for gene expression data without including additional patient data from other sources and focusing on classification problems was investigated in [9,10]. Partial Cox regression [11] is an extension of partial least squares to censored survival data; it has been proposed to analyze gene expression data. A modification that is less sensitive to outliers was proposed by [12]. Supervised principal component analysis only uses features that are correlated with survival time when computing principal components [13]. In "pre-conditioning" [14], supervised principal component analysis is first used to obtain a denoised outcome variable, which subsequently replaces the actual outcome when fitting a survival model with embedded feature selection. Perry et al. [15] analyzed text data from medical records of pediatric patients. They proposed supervised Laplacian eigenmaps, which combines Laplacian eigenmaps with a supervised loss function. Random indexing for dimensionality reduction of electronic health records to predict adverse drug reactions was proposed in [16].

Finally, several authors implemented comparative studies of feature selection and feature extraction methods for survival analysis in the past. A comparison of penalized Cox models with focus on low-dimensional data was presented in [17,18], where the latter studied gradient boosting methods as well. Regarding applications of survival analysis for microarray data, Benner et al. [19], Ma et al. [20] compared penalized Cox models, Schumacher et al. [21] analyzed univariate feature selection, partial Cox regression, and the least absolute shrinkage and selection operator (LASSO), and van Wieringen et al. [22] studied the performance of penalized Cox models, partial Cox regression, ensemble methods, and supervised principal component analysis. De Bin et al. [23] investigated univariate feature selection, forward stepwise selection, the LASSO, and boosting when combining low-dimensional clinical data with high-dimensional omics data.

In contrast to [9,10,16,15], the focus in our work is on survival analysis rather than classification. Moreover, we will not consider the problem of survival analysis applied to data with more features than samples ($p \gg n$), which has been extensively studied in the context of microarray data already (see e.g. [21,22]). The work presented by De Bin et al. [23] is the closest to our work, because they explicitly considered heterogeneous data consisting of low-dimensional clinical predictors and high-dimensional gene