DSSR

Discussion Paper No. 79

In-hospital Mortality Prediction for Trauma Patients Using Cost-sensitive MedLDA

Haruya Ishizuka, Tsukasa Ishigaki,
Naoya Kobayashi, Daisuke Kudo,
and Atsuhiro Nakagawa

March , 2018

Data Science and Service Research
Discussion Paper

Center for Data Science and Service Research
Graduate School of Economic and Management
Tohoku University
27-1 Kawauchi, Aobaku
Sendai 980-8576, JAPAN

In-hospital Mortality Prediction for Trauma Patients Using Cost-sensitive MedLDA

Haruya Ishizuka[1], Tsukasa Ishigaki[2], Naoya Kobayashi[3], Daisuke Kudo[3], and Atsuhiro Nakagawa[3]

[1] Graduate School of Economics and Management, Tohoku University, Miyagi, JAPAN isizuka19931114@gmail.com**
[2] Graduate School of Economics and Management, Tohoku University, Miyagi, JAPAN
[3] Tohoku University School of Medicine, Miyagi, JAPAN

Abstract. In intensive care units (ICUs), mortality prediction using vital sign or demographics of patients yields helpful information to support the decision-making of intensivists. Clinical texts recorded by medical staff tend to be valuable for prediction. However, text data are not applicable to outcome prediction of the regression framework in a direct way. In addition, learning of prediction models of such outcomes is a class of imbalanced data problem because the number of survivors is greater than the number of dead patients in most ICUs. To address these difficulties, we present Cost-Sensitive MedLDA: a supervised topic model employing cost-sensitive learning. The model realizes a prediction model from heterogeneous data such as vital signs, demographic information, and clinical text in an imbalanced class problem. Through experimentation and discussion, we demonstrate that the model has two benefits for use in medical fields: 1) our model has high prediction performance for minority instances while maintaining good performance for majority instances even if the training set is imbalanced data; 2) our model can reveal some characteristics that are associated with bad outcomes from the use of clinical texts.

1 Introduction

Intensive Care Units (ICUs) are closely associated with patient mortality. Intensivists must make prompt and accurate decisions to provide adequate treatment. Nevertheless, doing so manually imposes heavy burdens on them because of constraints related to time and available personnel. To address this issue, clinical decision support systems (CDSSs) able to detect critically ill patients have attracted attention. Accurate mortality prediction methods are fundamentally important for realizing practical CDSSs. Severity scores that are calculated from numerical data recorded in structured relational databases, have been used to evaluate illness severity. For instance, Acute Physiology and Chronic Health II

** The first author currently is an employee of a private company, however, the work was done at Tohoku University

(APACHE II)[7] and Simplified Acute Physiology Score II (SAPS II)[14] (Le Gall et al., 1993) are widely used at many hospitals.

In addition to numerical data, some researchers have recently developed mortality prediction methods using unstructured clinical text, and have reported its results [22, 8, 3, 16]. However, their methods are based mainly on a two-stage mortality prediction procedure. In this two-stage procedure, we first extract topics that represent underlying semantic themes of documents, from documents using an unsupervised topic model, e.g. Latent Dirichlet Allocation (LDA)[1]. In addition, we use the learned topic distribution as an input to a classifier. Some earlier works have demonstrated that supervised topic models (STMs), which can extract more predictive topics using label information associated with documents, outperform two-stage procedures in terms of classification accuracy[2, 11]. It is preferred to use STM for more accurate outcome prediction, but two main problems in earlier STMs arise when we use them for mortality prediction.

First, it is necessary to develop a cost-sensitive model. The imbalanced data problem is a well-known difficulty that poses challenges to the development and use of machine learning applications. Generally, in-hospital mortality rates in ICUs are from 10% to 15%, representing one example of imbalanced data. Learning algorithms for a classifier typically assume that instances in each class are distributed uniformly. When the training set comprises imbalanced data, a classifier rarely identifies minority instances. Consequently, the predictive accuracy becomes poor. Cost-sensitive learning is one approach to mitigating this difficulty. In cost-sensitive learning, if instances that actually belong to a minority class are misclassified to a majority class, a severer penalty is assigned to the objective function. However, no STM takes a cost-sensitive approach in its learning stage.

Secondly, for ICU settings, it is helpful to develop an STM that integrates textual information and numerical data. However, STMs have been developed mainly in the fields of document classification and image categorization. Presumably, topic representations and their discriminant functions are trained only from text or images. Electronic health records at an ICU include not only clinical text but also various numerical data. This information can be beneficial for outcome prediction.

To address these issue, we propose Cost-Sensitive Maximum entropy discriminative LDA (CS-MedLDA). Maximum entropy discriminative LDA (MedLDA) (Zhu et al., 2012[24]) is a supervised topic model that integrates discriminative max-margin learning with topic models. CS-MedLDA is its extension, which adopts a cost-sensitive approach in the training phase. In addition, our model can use numerical data as features of discriminant function, unlike conventional STM. As described herein, we demonstrate that the method can detect minority instances correctly and that it can discover characteristics that are associated with a minority group in an imbalanced dataset.

This paper is organized as follows. Sect.2 presents a summary of some related work. Sect.3 presents a description of the formulation and learning algorithm. Sect.4 explains the pre-experiment: one-versus-rest classification in 20

newsgroups dataset. Sect.5 presents experimentally obtained results in a trauma patient dataset. Finally, Sect.6 presents the clinical significance, conclusions, and avenues of future work.

2 Related Work

2.1 Cost-sensitive learning

Many works have specifically examined cost-sensitive learning[10]. Especially, the application of Support Vector Machine (SVM), a classifier based on the max-margin principle, has been investigated actively in this field because of its rigorous mathematical foundation and predictive power. As a well-known SVM application, cost-sensitive SVM (CS-SVM) [19], Boosting SVM[20], and Optimized cost-sensitive SVM[5] can be cited.

2.2 Supervised Topic Models

STMs, which are topic models that can learn predictive topic structures using label information of documents, have been developed mainly for document classification and image categorization[2, 15, 24]. Specifically, MedLDA has shown great promise for document classification, image categorization and link prediction in social networks[17, 21, 4]. In addition, some authors have developed sophisticated learning algorithms to reduce the training time of MedLDA[12, 25].

However, STMs are generally assumed as models that are trained with a single data type. In addition, no supervised topic model exists which uses cost-sensitive learning to manage an imbalanced data problem.

2.3 Mortality Prediction using LDA and Its Extension

Outcome prediction methods using textual information have been developed recently because of the advancement of Natural Language Processing techniques and text mining. Most of them use LDA and its extension to extract latent topics from clinical text[16, 8, 3, 22]. All of them reported the effectiveness of using clinical text and LDA variants.

However, these methods have used 2-stage procedure that learns topic representations and classifiers separately. Halpern et al. (2012) claim that STM has a higher prediction performance than two-stage procedures when the dimension of latent topic dimensions are low. Zhang et al.,2017[23] proposes a survival topic model which is an STM to estimate the hazard function of mortality. It is an example of using STM for mortality prediction. However, few works have explained the use of STM for mortality prediction. No report of the relevant literature describes the use of MedLDA to predict outcomes in spite of its superior classification ability.

3 Cost-sensitive MedLDA

This section presents the formulation and learning algorithm of CS-MedLDA. First, we introduce LDA, Regularized Bayesian Inference, and MedLDA as preliminaries. Next, we explain details of CS-MedLDA. In the following section, we respectively denote $\mathbf{W} = \{\mathbf{w}_d\}_{d=1}^D$ and $\mathbf{y} = \{y_d\}_{d=1}^D$ as the training documents and documents label. Here, $y_d \in \mathcal{Y} = \{-1, 1\}$. $KL[q\|p]$ stands for the Kullback–Leibler divergence. Also, $E_q[\cdot]$ represents the expectation.

3.1 Preliminaries

LDA LDA is a hierarchical Bayesian model that posits each document as an admixture of K topics, where each topic $\boldsymbol{\Phi}_k$ is a multinomial distributions over V vocabularies. For document d, the generating process can be described as presented below.

1. Draw a topic proportion $\quad \boldsymbol{\theta}_d \sim \mathrm{Dir}(\boldsymbol{\alpha})$.
2. For each word n $(0 \leq n \leq N_d)$
   (a) draw a topic assignment $z_{dn} \sim \mathrm{Mult}(\boldsymbol{\theta_d})$.
   (b) draw the observed word $w_{dn} \sim \mathrm{Mult}(\boldsymbol{\Phi}_{z_{dn}})$.

Therein, $\mathrm{Dir}(\cdot)$ represents a Dirichlet distribution, $\mathrm{Mult}(\cdot)$ denotes a multinomial, and $\boldsymbol{\Phi}_{z_{dn}}$ expresses the topic selected by the topic assignment z_{dn}. For Bayesian LDA, the topics are random samples drawn from a Dirichlet prior $\forall k,\ \boldsymbol{\phi}_k \sim \mathrm{Dir}(\boldsymbol{\beta})$.

Given a training set of documents $\mathbf{W}$, we let $\mathbf{z}_d = \{z_{dn}\}_{n=1}^{N_d}$ denote the set of topic assignments for document d and let $\mathbf{Z} = \{\mathbf{z}_d\}_{d=1}^D$ and $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_d\}_{d=1}^D$ respectively denote all the topic assignments and mixing proportions for the whole corpus. The goal of LDA is to infer the posterior distribution as

$$p(\boldsymbol{\Theta}, \boldsymbol{\Phi}, \mathbf{Z}|\mathbf{W}) = \frac{p_0(\boldsymbol{\Theta}, \boldsymbol{\Phi}, \mathbf{Z}) p(\mathbf{W}|\mathbf{Z}, \boldsymbol{\Phi})}{p(\mathbf{W})}, \tag{1}$$

where $p_0(\boldsymbol{\Theta}, \boldsymbol{\Phi}, \mathbf{Z}) = \prod_k p(\boldsymbol{\Phi}_k|\boldsymbol{\beta}) \prod_d p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) \prod_d \prod_n p(z_{dn}|\boldsymbol{\theta}_d)$,$p(\mathbf{W}|\mathbf{Z}, \boldsymbol{\Phi}) = \prod_d \prod_n p(w_{dn}|z_{dn}, \boldsymbol{\Phi})$ according to the generation process.

Regularized Bayesian Inference Jiang et al. (2012) show that (1) inferred from the Bayes rule is equivalent to the solution of the optimization problem on probability distribution below $q(\boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{\Phi})$ as

$$\min_{q(\mathbf{Z}, \boldsymbol{\Theta}, \boldsymbol{\Phi}) \in \mathbb{P}} KL[q(\mathbf{Z}, \boldsymbol{\Theta}, \boldsymbol{\Phi}) \| p_0(\mathbf{Z}, \boldsymbol{\Theta}, \boldsymbol{\Phi})] - E_q[\log p(\mathbf{W}|\mathbf{Z}, \boldsymbol{\Phi})],$$

where $\mathbb{P}$ signifies the space of probability distributions. Regularized Bayesian Inference (RegBayes) is the framework which regards posterior inference by Bayes rule as optimization problem including objective function and constraint on probability distribution. One benefit of this framework is that it can be extended naturally to include some regularization terms on q.

MedLDA As a supervised topic model, MedLDA learns topic representations and max-margin classifiers jointly. Training is conducted by solving the RegBayes problem below as

$$\begin{aligned} &\min_{q(\boldsymbol{\eta},\mathbf{Z},\boldsymbol{\Theta},\boldsymbol{\Phi})\in\mathbb{P}} \mathcal{L}(q(\boldsymbol{\eta},\mathbf{Z},\boldsymbol{\Theta},\boldsymbol{\Phi})) - 2C\mathcal{R}(q(\boldsymbol{\eta},\mathbf{Z},\boldsymbol{\Theta},\boldsymbol{\Phi})) \\ &\text{subject to } \forall d, 1 - y_d E_q[\boldsymbol{\eta}^{\mathrm{T}}\bar{\mathbf{z}_d}] - \xi_d \leq 0, \xi_d \geq 0 \end{aligned} \tag{2}$$

where $\boldsymbol{\eta}$ represents the coefficient vector of classifiers $\mathcal{L}(q)$ is $KL[q(\boldsymbol{\eta},\mathbf{Z},\boldsymbol{\Theta},\boldsymbol{\Phi})\|p(\boldsymbol{\eta},\mathbf{Z},\boldsymbol{\Theta},\boldsymbol{\Phi}|\mathbf{W})]$; $\mathcal{R}(q) = \sum_{d=1}^{D} E_q[\max(0, 1 - y_d\boldsymbol{\eta}^{\mathrm{T}}\bar{\mathbf{z}}_d)]$ denotes the expected hinge loss, and C signifies the penalty for misclassification. Also, $\bar{\mathbf{z}}$ is the average topic assignment vector with each element being $\bar{z}_k = \frac{1}{N}\sum_{n=1}^{N}\mathbb{I}(z_n = k)$. The label prediction rule is given as the signature function below:

$$sign(f(\boldsymbol{\eta}^{\mathrm{T}}\bar{\mathbf{z}})).$$

3.2 Cost-sensitive MedLDA

Formulation CS-MedLDA is a MedLDA variant that adopts cost-sensitive learning to learn topic representations and classifiers. In addition, this model can use not only documents but also numerical data, unlike previous supervised topic models. We define $\mathbf{s}_d,\ \mathbf{s}_d \in \mathbb{R}^P$ as numerical data. Let $\mathbf{x}_d = \{\bar{\mathbf{z}}_d, \mathbf{s}_d\}_{d=1}^{D}$ and $\boldsymbol{\eta} = \{\boldsymbol{\eta}_1, \boldsymbol{\eta}_2\},\ \boldsymbol{\eta}_1 \in \mathbb{R}^K,\ \boldsymbol{\eta}_2 \in \mathbb{R}^P$ respectively denote the feature vector and coefficient vector. Then, the training of CS-MedLDA is formulated with the optimization problem below as

$$\begin{aligned} &\min_{q(\boldsymbol{\eta},\mathbf{Z},\boldsymbol{\Theta},\boldsymbol{\Phi})\in\mathbb{P}} \mathcal{L}(q(\boldsymbol{\eta},\mathbf{Z},\boldsymbol{\Theta},\boldsymbol{\Phi})) - 2C^{+}\mathcal{R}^{+}(q(\boldsymbol{\eta},\mathbf{Z},\boldsymbol{\Theta},\boldsymbol{\Phi})) \\ &\qquad\qquad - 2C^{-}\mathcal{R}^{-}(q(\boldsymbol{\eta},\mathbf{Z},\boldsymbol{\Theta},\boldsymbol{\Phi})) \\ &\text{subject to } \forall d, 1 - y_d E_q[\boldsymbol{\eta}^{\mathrm{T}}\mathbf{x}_d] - \xi_d \leq 0, \xi_d \geq 0 \end{aligned} \tag{3}$$

where

$$\begin{aligned} \mathcal{R}^{+}(q) &= \sum_{d:\{d\in y_d=1\}} E_q[\max(0, 1 - y_d\boldsymbol{\eta}^{\mathrm{T}}\mathbf{x}_d)]; \\ \mathcal{R}^{-}(q) &= \sum_{d:\{d\in y_d=-1\}} E_q[\max(0, 1 - y_d\boldsymbol{\eta}^{\mathrm{T}}\mathbf{x}_d)]. \end{aligned}$$

The objective function is penalized with C^{+} if a minority instance is misclassified to a majority group. One can mitigate the imbalanced data problem by adjusting C^{-} and C^{+}. Here, CS-MedLDA is equivalent to MedLDA if $C^{+} = C^{-}$ and $\mathbf{s}_d = \mathbf{0}$.

Learning Algorithm We use data augmentation and Gibbs sampling[25] as a learning algorithm of CS-MedLDA. Furthermore, we designate $\mathbf{y}^{+}, \mathbf{W}^{+}$ and

$\mathbf{X}^+ = \{\mathbf{Z}^+, \mathbf{S}^+\}$ as the set of variables associated with a minority class, letting $\mathbf{y}^-, \mathbf{W}^-$ and $\mathbf{X}^- = \{\mathbf{Z}^-, \mathbf{S}^-\}$ b variables belonging to a majority class. Introducing the unnormalized pseudo-likelihood as

$$\phi(\mathbf{y}^+|\mathbf{X}^+, C^+, \boldsymbol{\eta}) = \prod_{d\in\{d:y_d=1\}} \exp(-2C^+ \max\{0, \xi_d\}),$$
$$\phi(\mathbf{y}^-|\mathbf{X}^-, C^-, \boldsymbol{\eta}) = \prod_{d\in\{d:y_d=-1\}} \exp(-2C^- \max\{0, \xi_d\}).$$

Then, the optimization problem (3) is rewritten as shown below:

$$\begin{aligned}\min_{q(\boldsymbol{\eta},\mathbf{Z},\boldsymbol{\Theta},\boldsymbol{\Phi})\in\mathbb{P}} \mathcal{L}(q(\boldsymbol{\eta},\mathbf{Z},\boldsymbol{\Theta},\boldsymbol{\Phi})) - E_q[\log\phi(\mathbf{y}^+|\mathbf{X}^+, C^+, \boldsymbol{\eta})] \\ - E_q[\log\phi(\mathbf{y}^-|\mathbf{X}^-, C^-, \boldsymbol{\eta})]\end{aligned}. \tag{4}$$

Using the ideas of data augmentation[18], the unnormalized pseudo-likelihood for a minority group can be expressed as

$$\begin{aligned}\phi(\mathbf{y}^+|\mathbf{X}^+, C^+, \boldsymbol{\eta}) &= \prod_{d\in\{d:y_d=1\}} \int_0^\infty \phi(y_d, \lambda_d|\mathbf{x}_d, C^+, \boldsymbol{\eta}) d\lambda_d \\ &= \prod_{d\in\{d:y_d=1\}} \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_d}} \exp\left(-\frac{(\lambda_d + C^+\xi_d)^2}{2\lambda_d}\right) d\lambda_d\end{aligned} \tag{5}$$

Then, if $y_d = 1$,

$$\phi(y_d, \lambda_d|\mathbf{x}_d, C^+, \boldsymbol{\eta}) = \frac{1}{\sqrt{2\pi\lambda_d}} \exp\left(-\frac{(\lambda_d + C^+\xi_d)^2}{2\lambda_d}\right).$$

In the case of $y_d = -1$, we obtain $\phi(y_d, \lambda_d|\mathbf{x}_d, C^-, \boldsymbol{\eta})$ by changing C^+ to C^-. Using $\phi(\mathbf{y}^+, \boldsymbol{\lambda}^+|\mathbf{X}^+, C^+, \boldsymbol{\eta})$ and $\phi(\mathbf{y}^-, \boldsymbol{\lambda}^-|\mathbf{X}^-, C^-, \boldsymbol{\eta})$, we can rewrite (3) as a new optimization problem including augmented variable $\boldsymbol{\lambda}$ as shown below.

$$\begin{aligned}\min_{q(\boldsymbol{\eta},\mathbf{Z},\boldsymbol{\Theta},\boldsymbol{\Phi})} \mathcal{L}(q(\boldsymbol{\eta},\mathbf{Z},\boldsymbol{\Theta},\boldsymbol{\Phi}) - E_q[\log\phi(\mathbf{y}^+, \boldsymbol{\lambda}^+|\mathbf{X}^+, C^+, \boldsymbol{\eta})] \\ - E_q[\log\phi(\mathbf{y}^-, \boldsymbol{\lambda}^-|\mathbf{X}^-, C^-, \boldsymbol{\eta})]\end{aligned} \tag{6}$$

Solving this optimization problem, we obtain the complete posterior as

$$q(\boldsymbol{\eta}, \boldsymbol{\lambda}, \mathbf{Z}, \boldsymbol{\Theta}, \boldsymbol{\Phi}) = \frac{p_0(\boldsymbol{\eta}, \mathbf{Z}, \boldsymbol{\Theta}, \boldsymbol{\Phi}) p(\mathbf{W}|\mathbf{Z}, \boldsymbol{\Phi}) \phi(\mathbf{y}^+, \boldsymbol{\lambda}^+|\mathbf{X}^+, C^+, \boldsymbol{\eta}) \phi(\mathbf{y}^-, \boldsymbol{\lambda}^-|\mathbf{X}^-, C^-, \boldsymbol{\eta})}{\psi(\mathbf{y},\mathbf{W})}, \tag{7}$$

where $\psi(\mathbf{y},\mathbf{W})$ is the normalized constant. In addition, we employ collapsed Gibbs sampling[9] by setting prior $p_0(\boldsymbol{\eta}, \mathbf{Z}, \boldsymbol{\Theta}, \boldsymbol{\Phi}) = p(\boldsymbol{\eta}) p_0(\mathbf{Z}, \boldsymbol{\Theta}, \boldsymbol{\Phi})$ and integrating out $\boldsymbol{\Theta}, \boldsymbol{\Phi}$ from $q(\boldsymbol{\eta}, \boldsymbol{\lambda}, \mathbf{Z}, \boldsymbol{\Theta}, \boldsymbol{\Phi})$.

Letting $q^*(\boldsymbol{\eta}|\text{rest})$, $q^*(z_{dn} = k|\text{rest})$ and $q^*(\lambda_d^{-1}|\text{rest})$, and assuming the full conditional distribution of $\boldsymbol{\eta}$, z_{dn} and λ_d^{-1}, these probability distributions are

Algorithm 1 Gibbs sampling for CS-MedLDA.

Initialization: set $\boldsymbol{\lambda} = \mathbf{1}$ and randomly draw z_{dn} from a uniform distribution
for $m = 1$ to M **do**
  draw the classifier $\boldsymbol{\eta}$ from the normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  for $d = 1$ to D **do**
    for each word n in document d **do**
      draw a topic from the multinomial distribution $M(C^+)$ or $M(C^-)$
    end for
    draw λ_d^{-1} (and thus λ_d) from the inverse Gaussian distribution $IG(1/C^+|\xi_d|, 1)$ or $IG(1/C^-|\xi_d|, 1)$
  end for
end for

given in the form presented below. Derivation details are presented in Appendix 1.

$$
\begin{aligned}
&q^*(\boldsymbol{\eta}|\text{rest}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&\boldsymbol{\Sigma} = \left(\frac{\mathbf{I}_{K+P}}{\nu^2} + (C^+)^2 \sum_{d \in \{d:y_d=1\}} \frac{\mathbf{x}_d \mathbf{x}_d^{\mathrm{T}}}{\lambda_d} + (C^-)^2 \sum_{d \in \{d:y_d=-1\}} \frac{\mathbf{x}_d \mathbf{x}_d^{\mathrm{T}}}{\lambda_d} \right)^{-1} \\
&\boldsymbol{\mu} = \boldsymbol{\Sigma} \left(C^+ \sum_{d \in \{d:y_d=-1\}} \frac{y_d(\lambda_d + C^+)}{\lambda_d} \mathbf{x}_d + C^- \sum_{d \in \{d:y_d=-1\}} \frac{y_d(\lambda_d + C^-)}{\lambda_d} \mathbf{x}_d \right)
\end{aligned}
\tag{8}
$$

$$
\begin{aligned}
q^*(z_{dn} = k|\text{rest}) \propto & \frac{(C_{k.-n}^t + \beta_t)(C_{d.-n}^k + \alpha_k)}{\sum_{t=1}^V C_{k.-n}^t + \sum_{t=1}^V \beta_t} \\
& \times \exp\left[\frac{\gamma \eta_{1k}^2 \{\lambda_d y_d + C^+(y_d - E_q[\boldsymbol{\eta}_2]^{\mathrm{T}} \mathbf{s}_d)\}}{\lambda_d} \right. \\
& \left. - (C^+)^2 \frac{\gamma^2 \eta_{1k}^2 + 2\gamma(1-\gamma)\eta_{1k} \Lambda_{dn}^k}{2\lambda_d} \right] \\
& = M(C^+)
\end{aligned}
\tag{9}
$$

If $y_d = -1$, we obtain $M(C^-)$ with changing C^+ to C^-. Table 1 shows the definition of $C_{d.-n}^k$,$C_{k.-n}^t$,γ and Λ_{dn}^k.

$$
q^*(\lambda_d^{-1}|\text{rest}) = IG(1/C^+|\xi_d|, 1) \tag{10}
$$

Here, $IG(a, b)$ is the inverse Gaussian distribution. After sampling λ_d^{-1}, we obtain λ_d by inversion. If $y_d = -1$, then in the same fashion as z_{dn}, we obtain $IG(1/C^-|\xi_d|, 1)$ with changing C^+ to C^-.

Algorithm 1 shows pseudo-code for the learning algorithm of CS-MedLDA. In our research, we set $\boldsymbol{\lambda} = \mathbf{1}$ initially and draw $\mathbf{Z}$ randomly from a uniform distribution.

Table 1. Definition of $C^k_{d.-n}$,$C^t_{k.-n}$,γ and Λ^k_{dn}

$C^k_{d.-n}$	$\sum_{n=1}^{N_d} \mathbb{I}(z_{dn}=k) - \mathbb{I}(z_{dn}=k)$
$C^t_{k.-n}$	$\sum_{d=1}^{D}\sum_{n=1}^{N_d} \mathbb{I}(z_{dn}=k)\mathbb{I}(w_{dn}=t) - \mathbb{I}(z_{dn}=k)\mathbb{I}(w_{dn}=t)$
γ	$\frac{1}{N_d}$
Λ^k_{dn}	$\frac{1}{N_d-1}\sum_{k'=1}^{K}\eta_{k'}C^{k'}_{d.-n}$

Table 2. Distribution of document labels in our experimental settings

Condition	$\lvert D^+\rvert$	$\lvert D^-\rvert$	*imbalance ratio*
1	798	4,852	14.1%
2	798	5,816	12.1%
3	798	6,803	10.4%
4	798	7,796	9.1%
5	798	8,787	8.3%

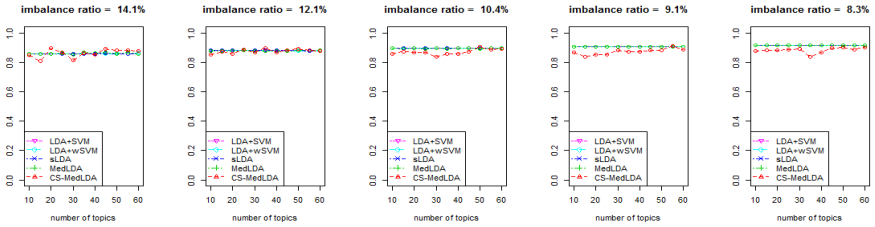
4 Pre-Experiment

4.1 Experimental Settings

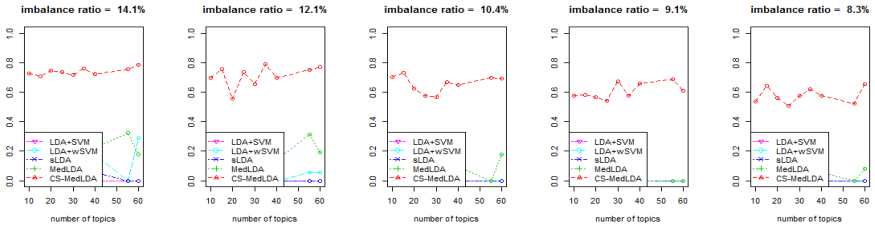
Sect.4 presents the pre-experiment results: one-versus-rest classification in a 20 newsgroup dataset. This dataset includes about 20,000 articles within 20 newsgroup categories. If an article d is labeled as *alt.athism*, then $y_d = 1$, else $y_d = -1$. In one-versus-rest classification, an imbalanced data problem occurs when numerous categories area used. Table 2 presents the distribution of labels in our settings. Column $|D^+|$ is the number of articles categorized into *alt.atishm*. Column $|D^-|$ is the number of articles with other labels. Column *imbalance ratio* is the ratio of *alt.athism* articles to the whole dataset. The training set contains 60% of total documents. The test set is the rest.

We compare the proposed model with four competitors: LDA+SVM, LDA+wSVM, sLDA, and MedLDA. Here, LDA+SVM and LDA+wSVM use LDA to learn the topic representations, and predict labels, respectively, with linear SVM and linear CS-SVM. sLDA is an STM proposed by Blei & McAliffe, 2007. sLDA and MedLDA are trained with Gibbs sampling[26, 25]. For LDA+SVM, sLDA and MedLDA, we set penalty parameter $C = 1$. For LDA+wSVM and CS-MedLDA, we penalize their objective function with penalty $C^+ = 2$ when a minority instance is misclassified to a majority group. In the opposite case, penalizing with $C^- = 1$. The $\boldsymbol{\eta}$'s prior $p_(\boldsymbol{\eta})$ for sLDA, MedLDA, and CS-MedLDA is the normal distribution $N(\mathbf{0}, \mathbf{I}_{K+P})$. For all five models, we use the symmetric Dirichlet prior $\boldsymbol{\alpha} = \mathbf{1}/K$ and $\boldsymbol{\beta} = \mathbf{1}/V, V = 61,118$. The number of MCMC iterations M is 100. The number of latent topics K is shifted from 10 to 60 in steps of 5.

To evaluate the predictive power, we use accuracy and the G-mean. The G-mean is the geometric mean of the accuracy of each class, i.e., G-mean $= \sqrt{\frac{\#TN}{|D^-|} \cdot \frac{\#TP}{|D^+|}}$, where the sizes of different classes have already been consid-



(a) Accuracy



(b) G-mean

Fig. 1. (a) Accuracy and (b) G-mean of each model when conducting one-versus-rest classification with a 20 newsgroup dataset. The x-axis shows the number of topics. The y-axis shows the value of indices. The *imbalance ratio* is shifted from 14.1% to 8.3%.

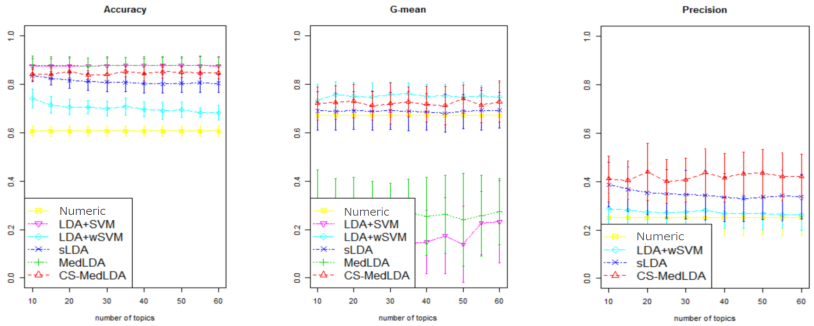
ered. Therefore, it is a good candidate for evaluating class-imbalance learning performance.

4.2 Result

Figure 1 presents the accuracy and G-mean results obtained using different methods. All models except CS-MedLDA show a low G-mean instead of high accuracy. This result indicates that these models cannot discriminate minority instances from majority instances because of the influence of an imbalanced data problem. However, the proposed model shows much higher G-mean. Moreover, its accuracy almost equals that of the others. Therefore, CS-MedLDA can more correctly identify *alt.athism* articles belonging to a minority group than competitors can.

5 Experimental Results in Trauma Patient Dataset

This section presents the results of in-hospital mortality prediction and topic evaluation for trauma patients in the Medical Information Mart for the Intensive Care-III (MIMIC-III) database[13]. MIMIC-III includes numerical data and clinical text of patients who were admitted to the Beth Israel Deaconess Medical Center during 2001–2012. We extract 1,546 trauma patients (in-hospital mortality rate: 12.4%) who were older than 16 years old and who had stayed in this ICU for 24 hr or longer.



(a) Accuracy (b) G-mean (c) Precision

Fig. 2. (a) Accuracy, (b) G-mean, and (c) Precision of each model when predicting patient outcomes in a trauma patient dataset.

The material documents are nursing notes and radiology reports gathered within the first 24 hr of the ICU stay. As preprocessing, we first eliminate punctuation, numbers, and clinical stopwords[4]. Next, we extracted noun phrases using morphological analysis. Finally, we obtained 18,041 documents and 28,080 vocabularies from 1,546 trauma patients.

Additionally, we obtained 12 numerical data such as gender, heart rate, and hematocrit. All of these data are used to calculate SAPS II or APACHE II, intensivists normally use in their work.

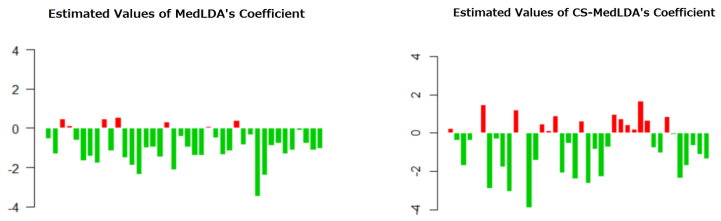
5.1 In-hospital Mortality Prediction

This experiment investigated the in-hospital mortality prediction accuracy of the proposed model. We compare the proposed model with five competitors: Numeric, LDA+SVM, LDA+wSVM, sLDA, and MedLDA. The numeric model predicts patient outcomes using linear SVM, but this model is based only on numerical data without text. Other models are the same as those described in Sect.4 except that models predict outcomes using both numerical and textual information. Hyperparameters are also the same settings as those described in Sect.4.

For evaluating predictive capability, we use precision in addition to accuracy and the G-mean. Precision is defined as $\text{precision} = \dfrac{\#TP}{\#TP + \#FP}$, indicating the percentage of false alerts.

Results Fig. 2 presents results of this experiment. MedLDA and LDA+wSVM show the highest accuracy of the six models, although their G-means are much lower than those of other competitors. These models cannot detect critically

[4] `https://github.com/kavgan/clinical-concepts/blob/master/clinical-stopwords.txt`



(a) MedLDA (b) CS-MedLDA

Fig. 3. (a) is the Monte-Carlo estimates of $\boldsymbol{\eta}_1$ of original MedLDA, and (b) is CS-MedLDA's one when the number of topics $K = 40$.

ill patients by the influence of an imbalanced data problem. When a classifier faces an imbalanced data problem, its precision becomes unstable because the denominator is extremely small. Therefore, we did not calculate their precision.

Numeric and LDA+wSVM show higher G-mean than either LDA+SVM or MedLDA. However, their accuracy and precision were lower than that of either sLDA or CS-MedLDA. This result indicates that these models classified instances which actually belong to a majority group to a minority group in surplus. Comparing CS-MedLDA with sLDA, the former shows higher values than sLDA with all indices. Therefore, our proposed model can correctly predict patient outcomes as compared to its competitors.

5.2 Topic Evaluation

This section presents the result of investigation into the topics extracted using CS-MedLDA. First, we explore the difference of coefficient vector $\boldsymbol{\eta}_1$ between standard MedLDA and our proposed model. Next, based on estimates of coefficients η_{1k} and frequently occurring words in each topic, we strive to discover topics that have a positive correlation with bad outcomes.

coefficient $\boldsymbol{\eta}$ Analysis Fig. 3 presents the estimated values of coefficient $\boldsymbol{\eta}_1$ of MedLDA and CS-MedLDA in the trauma patient dataset. We considered first 50 MCMC iterations as the burn-in period. The estimates of $\boldsymbol{\eta}_1$ of the original MedLDA tend to be negative values. Based on the results of two classification experiments, MedLDA produces poor predictions because it is unable to capture topics that are positively correlated with minority instances. However, CS-MedLDA can capture such topics by giving a larger penalty for misclassifying minority instances. As a result, CS-MedLDA produces better predictions.

Qualitative Evaluation Table 3 presents the most frequently observed words of 10 topics extracted using CS-MedLDA. The estimated value of η_{1k} corre-

sponding to each topic is positive. Topic 1 has the strongest positive correlation from the viewpoint of the value of η_{1k}.

Hematoma, scan and subarachnoid hemorrhage (sah) appear as frequent words of topic 1. Therefore, this topic is the result of CT scan over subarachnoid hemorrhage patients. In addition, dislocation of a brain might occur from the fact that **dislocation** appears frequently in this topic. Therefore, patients whose ratio of this topic is high might be afflicted by a severe subarachnoid hemorrhage.

In topic 3, malignant disease spreads to the subclavian lymph node: **lymph** and **clavicle** were observed frequently. In addition, patients were in **postop (post-operation) agitation**. Metastasis of malignant disease and post-operation agitation are known as risk factors related to bad outcomes.

Topic 4 represents tracheal intubation for mechanical ventilation. Actually, **ett** is an abbreviation of endotracheal tube. Generally, endotracheal tubes are used to secure a patient airway with consciousness disturbance or respiratory failure. Because **lobe** also appeared among the most frequently observed words, this topic represents respiratory care.

Both of these topics include factors related to deceased patients from past findings. From the above, CS-MedLDA can discover characteristics associated with patient mortality. However, some topics are difficult to interpret. Improvement of interpretability is an important task for future work.

6 Clinical Significance and Conclusion

As described herein, we proposed a CS-MedLDA, a supervised topic model using cost-sensitive learning to ascertain the latent topic structure and max-margin classifiers. Experimentation revealed that CS-MedLDA has two preferred properties. The proposed model detects minority instances more correctly than previous methods under an imbalanced data problem; it discovers some characteristics associated with a minority group from documents. These properties bring two benefits when applied in medical fields.

Table 3. Most frequently observed words in 10 topics with coefficient ($\eta \geq 0$).

Topic 1	hematoma,mm,sinus,optiray,dislocation,scan,npn,foci, opacities,car,sah
Topic 2	ct,year,number,radiology,head,spine,hemorrhage, status,resp,structures
Topic 3	process,checks,stool,lymph,tol,clavicle,multiplanar, reformats,need,postop ,agitation
Topic 4	lobe,line,namepattern,aorta,ett,opacification,site, endotracheal,elbow,ribs,vein
Topic 5	foley,sbp,side,pupils,gallbladder,eye,qhr, remains,mcgkgmin,diameter,tree
Topic 6	herniation,service,read,flow,extubation,scale,spaces, question,rule,sicu,ed
Topic 7	indication,blood,rib,lat,detail,foot,transfer,units, volumes,data,drain
Topic 8	wcontrast,atelectasis,bilat,ventricle, lymph,adenopathy,rt,arrival,breath, reduction,heights
Topic 9	urine,midline,bp,npo,nursing,collar,wean,pancreas, stenosis,displacement,tls
Topic 10	findings,lung,time,abd,neck,position,exam,leg, mdct,calcifications,hegiht

First, our model is expected to work well for other imbalanced data classification problems. In medical fields, various imbalance data are generated aside from patient mortality. One example is disease recurrence. In addition, numerical data and clinical text are recorded all over the hospital to monitor patients' disease state. Using the proposed model, we can readily construct a prediction model that integrates numerical and textual information with a simple Gibbs sampling algorithm. These two properties, i.e. detectability over minority instances and the ease of integrating numerical and textual information, are helpful for other imbalanced data classification.

Secondly, the proposed model might become a powerful method by which we can explore the characteristics associated with a minority group. In conventional clinical studies, researchers have attempted to seek such characteristics for preventing disease onset or finding care methods from numerical data. In other words, researchers have often been interested in the characteristics of minority instances in their studies because patients with bad outcomes or those afflicted with some disease are rare examples. Through interpretation of topics extracted using our model, we can uncover them not from numerical data but from unstructured text. Therefore, CS-MedLDA might become another tool to elucidate minority group characteristics.

Future studies will be conducted to exploit dynamic supervised topic models that learn topic representations and its dynamics from documents and from supervision. Patient conditions might change an instant after ICU admission. Specifically, one's health condition tends to be unstable in his/her acuity phase. Use of this dynamic information of clinical data has presented many benefits for outcome prediction. Dynamic supervised topic models might become an effective outcome prediction method based on dynamic information and supervision, and a text mining method that captures the time-variant evolution of illness severity from clinical texts.

A Derivation Detail of Conditional Distributions

By integrating out the Dirichlet variables $(\boldsymbol{\Theta}, \boldsymbol{\Phi})$ in the complete posterior distribution (7), we obtain the collapsed posterior distribution

$$\begin{aligned} q^*(\boldsymbol{\eta}, \boldsymbol{\lambda}, \mathbf{Z}) &\propto p(\boldsymbol{\eta}) p(\mathbf{W}, \mathbf{Z} | \boldsymbol{\alpha}, \boldsymbol{\beta}) \phi(\mathbf{y}^+, \boldsymbol{\lambda}^+ | \mathbf{X}^+, C^+, \boldsymbol{\eta}) \phi(\mathbf{y}^-, \boldsymbol{\lambda}^- | \mathbf{X}^-, C^-, \boldsymbol{\eta}) \\ &= p(\boldsymbol{\eta}) \left[\prod_{d=1}^{D} \frac{\boldsymbol{\delta}(\mathbf{C}_d + \boldsymbol{\alpha})}{\boldsymbol{\delta}(\boldsymbol{\alpha})} \right] \prod_{k=1}^{K} \frac{\boldsymbol{\delta}(\mathbf{C}_k + \boldsymbol{\beta})}{\boldsymbol{\delta}(\boldsymbol{\beta})} \\ &\times \prod_{d \in \{y_d = 1\}} \frac{1}{\sqrt{2\pi\lambda_d}} \exp\left(-\frac{(\lambda_d + C^+ \xi_d)^2}{2\lambda_d} \right) \\ &\times \prod_{d \in \{y_d = 1\}} \frac{1}{\sqrt{2\pi\lambda_d}} \exp\left(-\frac{(\lambda_d + C^- \xi_d)^2}{2\lambda_d} \right) \end{aligned} \tag{11}$$

where $\boldsymbol{\delta}(\mathbf{x}) = \frac{\prod_{i=1}^{dim(\mathbf{x})} \Gamma(x_i)}{\Gamma(\sum_{i=1}^{dim(\mathbf{x})} x_i)}$,$\Gamma(\cdot)$ is the Gamma function. Also, $\mathbf{C}_k = \{C_k^t\}_{t=1}^V$ is the set of word counts associated with topic k; $\mathbf{C}_d = \{C_d^k\}_{k=1}^K$ is the number of times that terms are associated with topic k. The full conditional distributions used in collapsed Gibbs sampling are the following.

For $\boldsymbol{\eta}$ Setting Gaussian prior $N(\mathbf{0}, \nu^2\mathbf{I}_{K+P})$ for $\boldsymbol{\eta}$, the conditional distribution $q^*(\boldsymbol{\eta}|\text{rest})$ is given by as the following Gaussian distribution.

$$
\begin{aligned}
q^*(\boldsymbol{\eta}|\text{rest}) &\propto p(\boldsymbol{\eta}) \exp\left(\sum_{d\in\{d:y_d=1\}} \frac{-(\lambda_d + C^+\xi_d)^2}{2\lambda_d} \right) \exp\left(\sum_{d\in\{d:y_d=-1\}} \frac{-(\lambda_d + C^+\xi_d)^2}{2\lambda_d} \right) \\
&\propto N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
\boldsymbol{\Sigma} &= \left(\frac{\mathbf{I}_{K+P}}{\nu^2} + (C^+)^2 \sum_{d\in\{d:y_d=1\}} \frac{\mathbf{x}_d\mathbf{x}_d^{\mathrm{T}}}{\lambda_d} + (C^-)^2 \sum_{d\in\{d:y_d=-1\}} \frac{\mathbf{x}_d\mathbf{x}_d^{\mathrm{T}}}{\lambda_d} \right)^{-1} \\
\boldsymbol{\mu} &= \boldsymbol{\Sigma}\left(C^+ \sum_{d\in\{d:y_d=-1\}} \frac{y_d(\lambda_d + C^+l)}{\lambda_d}\mathbf{x}_d + C^- \sum_{d\in\{d:y_d=-1\}} \frac{y_d(\lambda_d + C^-l)}{\lambda_d}\mathbf{x}_d \right)
\end{aligned}
\tag{12}
$$

Therefore, we can draw coefficient vector $\boldsymbol{\eta}$ from $q(\boldsymbol{\eta}|rest) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

For Z . The conditional distribution of $\mathbf{Z}$ given the other variable is the following.

$$
\begin{aligned}
q^*(\mathbf{Z}) \propto \prod_{k=1}^{K} \frac{\boldsymbol{\delta}(\mathbf{C}_k + \boldsymbol{\beta})}{\boldsymbol{\delta}(\boldsymbol{\beta})} \prod_{d\in\{y_d=1\}} \frac{\boldsymbol{\delta}(\mathbf{C}_d + \boldsymbol{\alpha})}{\boldsymbol{\delta}(\boldsymbol{\alpha})} \exp\left(-\frac{(\lambda_d + C^+\xi_d)^2}{2\lambda_d} \right) \\
\times \prod_{d\in\{y_d=-1\}} \frac{\boldsymbol{\delta}(\mathbf{C}_d + \boldsymbol{\alpha})}{\boldsymbol{\delta}(\boldsymbol{\alpha})} \exp\left(-\frac{(\lambda_d + C^-\xi_d)^2}{2\lambda_d} \right)
\end{aligned}
$$

By canceling common factors, we can derive the conditional distribution as $q^*(z_{dn}|\text{rest})$.

For $\boldsymbol{\lambda_d}$ Finally, the conditional distribution of the augmented variable $\boldsymbol{\lambda}$ given the rest variable is a generalized inverse Gaussian distribution (Devroye, 1986) as

$$
\begin{aligned}
q^*(\lambda_d|\text{rest}) &\propto \frac{1}{\sqrt{2\pi\lambda_d}} \exp\left(-\frac{(\lambda_d + C^+\xi_d)^2}{2\lambda_d} \right) \\
&\propto \frac{1}{\sqrt{2\pi\lambda_d}} \exp\left(-\frac{(C^+)^2\xi_d^2}{2\lambda_d} - \frac{\lambda_d}{2} \right) \\
&= GIG(\frac{1}{2}, 1, (C^+)^2\xi_d^2),
\end{aligned}
$$

where $GIG(p,a,b) = C(p,a,b)x^{p-1}\exp\left(-\frac{1}{2}\left(\frac{b}{x}+ax\right)\right)$ and $C(p,a,b)$ is a normalization constant. As $q^*(\lambda_d|\mathbf{X},\boldsymbol{\eta})$ is the generalized inverse Gaussian distribution, λ_d^{-1} follows an inverse Gaussian distribution. We denote $q^*(\lambda_d^{-1}|rest)$ as the conditional distribution of λ_d^{-1}, and $q^*(\lambda_d^{-1}|rest)$ is the following

$$q(\lambda_d^{-1}|rest) = IG(1/C^+|\xi_d|,1),$$

where $IG(a,b) = \sqrt{\frac{b}{2\pi x^3}}\exp\left(-\frac{b(x-a)^2}{2a^2x}\right), a,b>0.$

References

[1] D.M. Blei, A. Ng, and M.I. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research (JMLR), 3:993–1022, 2003.

[2] D.M. Blei and J.D. McAuliffe. Supervised topic models. In Advances in Neural Information Processing Systems (NIPS), pages 121–128, 2007.

[3] K.L. Caballero and R. Akella. Dynamic modelling patient's health state from electric medical records: A time series approach. Proc. 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.69-78, August, 2015.

[4] N. Chen, J. Zhu, F. Xia, and B. Zhang. Discriminative relational topic models. IEEE Trans. on PAMI, 37(5):973–986, 2015.

[5] P. Cao, D. Zhao, and O. Zaiane. An optimized cost-sensitive SVM for imbalanced data learning. In Advances in Knowledge Discovery and Data Mining: 2013. Berlin Heidelberg: Springer; 2013:28092.

[6] L. Devroye. Non-uniform random variate generation. Springer-Verlag, 1986.

[7] J.J. Escarce and M.A. Kelly. Admission source to the medical intensive care unit predicts hospital death independent of APACHE II score. Journal of American Medical Association, 264(18):.2389–2394, 1985.

[8] M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits. Unfolding physiological state: Mortality modelling in intensive care units, Proc. 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.75–84, 2014.

[9] T.L. Griffiths and M. Steyvers. Finding scientific topics. Proceedings of the National Academy of Science (PNAS), 5228–5235, 2004.

[10] H. He and E.A. Garcia. Learning from imbalanced data, IEEE Trans. Knowl. Data Eng., 21(9):1263–1284, 2009.

[11] Y. Halpern, S. Horng, L. Nathanson, N. Shapiro, and D. Sontag. A comparison of dimensionality reduction techniques for unstructured clinical text, In ICML 2012 Workshop on Clinical Data Analysis, 2012.

[12] Q. Jiang, J. Zhu, M. Sun, and E.P. Xing. Monte Carlo methods for maximum margin supervised topic models. Advances in Neural Information Processing Systems (NIPS), 1601–1609, 2012.

[13] A. Johnson, T. Pollard, L. Shen, L. Lehman, M. Fen, M. Ghassemi, B. Moody, P. Szolovits, L. Celi, and R. Mark. MIMIC-III a freely accessible critical care database. Scientific Data, 2016.

[14] L.G. JR, L.S, and S.F. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. Journal of American Medical Association, 270(24):2957–2963, 1993.
[15] S. Lacoste-Jullien, F. Sha, and M.I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. NIPS, 897–904, 2008.
[16] L. Lehman, M. Mohammed Saeed, W. Long, J. Lee, and R. Mark. Risk stratification of icu patients using topic models inferred from unstructured progress notes. AMIA Annual Symposium Proc., 2012:505–511, 2012.
[17] D. Li, S. Somasundaran, and A. Chakraborty. A combination of topic models with max-margin learning for relation detection. In ACL TextGraphs-6 Workshop, 2011.
[18] N.G. Polson and S.L Scott. Data augmentation for support vector machines. Bayesian Analysis, 6(1):1–4, 2011.
[19] K. Veropoulos, C. Campbell, and N. Cristianini. Controlling the sensitivity of support vector machines. In: Proceedings of the International Joint Conference on Artificial Intelligent, pp. 55–60, 1999.
[20] B.X. Wang and N. Japkowicz, Boosting support vector machines for imbalanced datasets. Journal of Knowledge and Information Systems, 4994:38–47, 2008.
[21] Y. Wang and G. Mori. Max-margin latent Dirichlet allocation for image classification and annotation. In BMVC, 2011.
[22] J. Yohan and C.P. Rose. Time Series Analysis of Nursing Notes for Mortality Prediction via a State Transition Topic Model. Proceedings of the 24th ACM International Conference on Information and Knowledge Management, 2015.
[23] Y. Zhang, R. Jiang, and L. Petzold, Survival Topic Models for Predicting Outcomes for Trauma Patients. Data Engineering (ICDE), 2017 IEEE 33rd International Conference.
[24] J. Zhu, A. Ahmed, and E.P. Xing. MedLDA: maximum margin supervised topic models. Journal of Machine Learning Research (JMLR), (13):2237–2278, 2012.
[25] J. Zhu, N. Chen, H. Perkins, and B. Zhang. Gibbs max-margin topic models with fast inference algorithms. In International Conference on Machine Learning (ICML), pp. 124–132, 2013.
[26] J. Zhu, X. Zheng, and B. Zhang. Bayesian logistic supervised topic models with data augmentation. In The Annual Meeting of the Association for Computational Linguistics, 2013.