Yu *et al. BMC Medical Informatics and Decision Making (2022) 22:14*
https://doi.org/10.1186/s12911-022-01747-3

BMC Medical Informatics and Decision Making

RESEARCH Open Access

Prediction and risk stratification from hospital discharge records based on Hierarchical sLDA

Guanglei Yu[1], Linlin Zhang[2], Ying Zhang[3], Jiaqi Zhou[1], Tao Zhang[1] and Xuehua Bi[1*]

Abstract

Background: The greatly accelerated development of information technology has conveniently provided adoption for risk stratification, which means more beneficial for both patients and clinicians. Risk stratification offers accurate individualized prevention and therapeutic decision making etc. Hospital discharge records (HDRs) routinely include accurate conclusions of diagnoses of the patients. For this reason, in this paper, we propose an improved model for risk stratification in a supervised fashion by exploring HDRs about coronary heart disease (CHD).

Methods: We introduced an improved four-layer supervised latent Dirichlet allocation (sLDA) approach called Hierarchical sLDA model, which categorized patient features in HDRs as patient feature-value pairs in one-hot way according to clinical guidelines for lab test of CHD. To address the data missing and imbalance problem, RFs and SMOTE methods are used respectively. After TF-IDF processing of datasets, variational Bayes expectation-maximization method and generalized linear model were used to recognize the latent clinical state of a patient, i.e., risk stratification, as well as to predict CHD. Accuracy, macro-F1, training and testing time performance were used to evaluate the performance of our model.

Results: According to the characteristics of our datasets, i.e., patient feature-value pairs, we construct a supervised topic model by adding one more Dirichlet distribution hyperparameter to sLDA. Compared with established supervised algorithm Multi-class sLDA model, we demonstrate that our proposed approach enhances training time by 59.74% and testing time by 25.58% but almost no loss of average prediction accuracy on our datasets.

Conclusions: A model for risk stratification and prediction of CHD based on sLDA model was proposed. Experimental results show that Hierarchical sLDA model we proposed is competitive in time performance and accuracy. Hierarchical processing of patient features can significantly improve the disadvantages of low efficiency and time-consuming Gibbs sampling of sLDA model.

Keywords: Risk stratification, Topic models, Supervised latent Dirichlet allocation, Hospital discharge records

Introduction

Cerebrovascular accident (CVA), coronary heart disease (CHD) and other cardiovascular diseases (CVD) are the leading causes of death and serious family burden in China nowadays. According to the World Health Organization (WHO), risk factors can increase the chances that a person suffers from that disease (WHO, 2014). Risk stratification incorporating these risk factors can be used by physicians to assess the risk of atherosclerotic of individual patient, such as taking treatment with drugs to lower blood pressure and blood cholesterol based on an individual's absolute cardiovascular risk [1]. According to

*Correspondence: ak422@163.com
[1] School of Medical Engineering and Technology, Xinjiang Medical University, No.567 North Shangde Road, Urumqi, China
Full list of author information is available at the end of the article

accurate risk stratification, so as to reduce the overall risk of CVD or CHD, physicians can formulate corresponding comprehensive clinical treatments or life intervention management programs for patients with different risk levels. To carry out risk stratification also plays an important role in individualized nursing, drug development and cost estimation of CVD [2].

In recent years, with the rapid development of medical information technology, the applications of electronic medical records (EMRs) are becoming more and more widely and in-depth. EMRs store and share clinical information of patients, such as general items, diagnostic images (e.g., X-rays), history of present illness (HPI), family history, lab results etc. for different diseases [3–5] and various medical applications [6–9], especially HDRs, which include abundant information on patient risk factors. In 2014, in order to identify and extract medical risk factors related to CHD, i2b2/UTHealth Natural Language Processing shared task with respect to the longitudinal medical records of patients with diabetes. The risk factors included hypertension, hyperlipidemia, obesity, smoking status, family history, diabetes [10].

Multivariable traditional assessment models have been provided to estimate CVD risk [11, 12]. At the cohort level, these risk stratification models take statistical analysis techniques (e.g., logistic regression, Cox regression, etc.) to estimate the absolute risk of patients, which offer little insight beyond a flat score-based segmentation that has high cost, carefully selected and highly stratified patient characteristics [13].

In this study, we proposed a four-layer probabilistic topic model, i.e., Hierarchical sLDA model, for risk stratifications and prediction of CHD by using the diagnosis cases from HDRs, where Hierarchical sLDA model is a variant of sLDA. Firstly, we collected data from real clinical settings and annotated risk factors with an annotation tool developed by ourselves, under the guidance of clinicians. Then, we extracted the patient feature-value pairs of risk factors and encoding them in One Hot way according to clinical guidelines for lab test of CHD. Meanwhile, to address the data missing and imbalance problem, RFs and SMOTE are used respectively. After TF-IDF processing of datasets, variational Bayes Expectation-maximization (VBEM) and generalized linear model (GLM) was used to recognize the latent clinical state of a patient, i.e., risk stratification, as well as to predict CHD. Experimental results show that our model can significantly improve the disadvantages of low efficiency and time-consuming of sLDA model.

Related work

With the continuous digitization and storage of knowledge in the forms of news, blogs, web pages, scientific articles, books, images, sound, video and social networks, it is more and more difficult for us to find what we are looking for from massive information [14].

In 1999, Thomas Hofmann proposed probabilistic latent semantic indexing (PLSI), which characterizes the polysemy of a word by describing the word frequency vector with multinomial distribution [15]. The proposal of PLSI enables the discovery and analysis of potential topics or categories in a large number of documents, and realizes the tasks of document clustering and dimension reduction.

In order to overcome the defects of inconsistent generative semantics of PLSI that cannot generate (i.e., predict) new documents and its model overfitting, Blei et al. [16] proposed LDA (latent Dirichlet allocation) topic model in 2003, and two commonly used approximate inference methods are Variational Bayes (VB) deterministic and collapsed Gibbs sampling (GS) stochastic approximation. Girolami and Kabán [17] showed that PLSI is maximum a posteriori (MAP) estimated LDA model under uniform Dirichlet prior. LDA is a three-layer hierarchical (including documents, topics and words) Bayesian unsupervised topic model. Based on the Bag-of-words (BOW) representation, LDA can cluster topics or classify texts from a large number of documents and has good scalability. With the development of probabilistic topic model, it has made continuous progress in image analysis, bioinformatics and other fields [18]. Jelodar et al. [19] reviewed the research progress, future development trend and wide application subjects of LDA topic model from 2003 to 2016, such as Social Network, Crime Science, Medical/Biomedical and Linguistic science.

For text classification, Li and McCallum [20] showed that LDA model does not capture the correlation between topics, and the accuracy and efficiency of topic classification are not outstanding and insufficient. Furthermore, although LDA model can achieve document clustering and dimension reduction and other tasks, it is not suitable for prediction. In 2010, Blei and McAuliffe [21] proposed supervised latent Dirichlet allocation (sLDA). By combining GLM for latent topics, and selecting different exponential distribution family according to response variables, multiple response variables (e.g., real, category or multinomial response variable) can be predicted.

Wang et al. [22] extended to image classification and proposed Multi-class sLDA prediction model by selecting GLM model for multinomial response. Their experimental results showed that the average accuracy of the model for LabelMe datasets (1600 images, 8 categories) and UIUC-Sport datasets (1792 images, 8 categories) was 76% and 66% respectively, which was better than that of Li and Perona [23] and Bosch et al. [24], and the average accuracy increased by more than 10%.

By using factor graph to represent collapsed LDA to encode the joint probability, Zeng et al. [25] enabled the belief propagation (BP) for approximate inference and parameter estimation and has enhanced both speed and accuracy by experimental results on four document datasets.

Besides these methods, there are various variants and applications about the state-of-the-art topic model LDA recently, e.g., opinion mining on big data [26], stock market returns [27], topic change point detection [28], open-ended versus closed-ended response [29], etc.

Motivated by these observations, our study proposes a probabilistic ensemble classification method, which distinguishes from other methods in that: (1) it takes a slight structural change to standard sLDA, which improves three-layer sLDA to a four-layer called Hierarchically sLDA model, and achieves encouraging experimental results in terms of time performance; (2) our model can provide prediction of CHD and risk stratification simultaneously.

Table 1 Symbols and notations

Symbols	Notations
$1 \leq d \leq D$	HDRs index
$f_{1:F}$	Patient features
$v_{1:N}$	Patient feature-value pairs
$1 \leq k \leq K$	Topic index
π_k	Topic-feature multinomials of feature *f*, π_k is *F*-dimensional vector
β_k	Topic-value multinomials of feature-value *v*, β_k is *N*-dimensional vector
α	*K*-dimensional Dirichlet parameter vector
θ	*K*-dimensional topic proportions
$r_{1:N}$	Topic assignments
y	Response variables
$\eta_{1:C}$	Class coefficients

The remaining sections of this paper are organized as follows. "Hierarchical sLDA" section describes our proposed model, variational inference and parameter estimation. "Experiments" section carefully describes the datasets and presents our model experimental results. Finally, we present our conclusions possible directions for future work in "Conclusions" section.

Hierarchical sLDA

Modeling HDRs and labels

Firstly, we summarize some important symbols and notations in this paper shown in Table 1.

The graphical model representation of hierarchical sLDA is depicted in Fig. 1. Nodes are random variables; edges indicate possible dependence; a shaded node is an observed variable; an unshaded node is a hidden variable.

Each HDRs is represented as a bag of patient feature $f_{1:F}$ or patient feature-value pair $v_{1:N}$. The category c is a discrete class label. Each topic is a distribution over a vocabulary of patient feature, and also be regarded as distribution over vocabulary of patient feature-value pair. K is the number of latent topics; N is the number of feature-value of a single HDRs; D is the number of patient HDRs.

Our model assumes the following generative process of an HDRs, and its class label.

1. Draw topic proportions $\theta, \theta \mid \alpha \sim \text{Dir}(\alpha)$;
2. For each patient HDRs feature-value pair $v_{1:N}$:

   (a) Draw topic assignment r_n from category distribution with parameter $\theta, r_{\text{n}} \mid \theta \sim \text{Mult}(\theta)$;
   (b) Draw feature f_n from category distribution with parameter π, $f_n \mid r_n, \pi_{1:K} \sim \text{Mult}\left(\pi_{r_n}\right)$;
   (c) Draw feature-value pair v_n from category distribution with parameter β, $v_n \mid r_n, f_n, \beta_{1:K} \sim \text{Mult}\left(\beta_{f_n, r_n}\right)$;

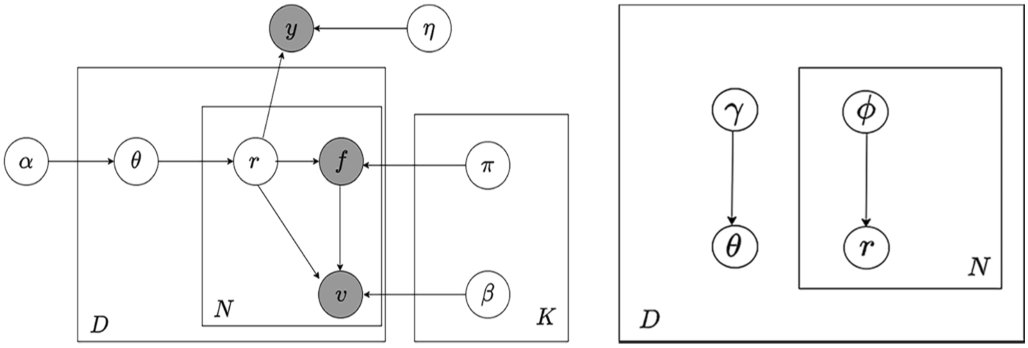


Fig. 1 Probabilistic graphical model. The probabilistic graphical model representation of Hierarchical sLDA (left); the graphical model representation of variational distribution (right)

3. Draw class label y from GLM distribution $y \mid r_{1:N}, \eta \sim GLM(\bar{r}, \eta)$, where $\bar{r} = \frac{1}{N}\sum_{n=1}^{N} r_n$, that is:

$$p\left(y \mid r_{1:N}, \eta\right) = \exp\left(\eta_y^T \bar{r}\right) / \sum_{l=1}^{C} \exp\left(\eta_l^T \bar{r}\right)$$

it also can be written as exponential distribution family:

$$p\left(y \mid r_{1:\mathrm{N}}, \eta\right) = \exp\left\{\eta_y^T \bar{r} - \log\left(\sum_{l=1}^{C} \exp\left(\eta_l^T \bar{r}\right)\right)\right\}$$

The posterior inference of the model can be divided into three steps or tasks.

(a) *Variational approximate inference* A posterior inference is to compute the conditional distribution of the latent variables at the patient HDRs level. That is, approximate inference is used to estimate the parameters of the variational distribution.
(b) *Parameter estimation* The model parameters $\alpha, \beta_{1:K}, \pi_{1:K}, \eta$ are fitted with variational expectation maximization (EM) by maximum likelihood estimation.
(c) *Prediction and risk stratification* To perform prediction and risk stratification over the model parameters$\alpha, \beta_{1:K}, \pi_{1:K}, \eta$ and variational distribution parameters γ, ϕ means approximation of posterior expectation of response variable $y = \mathrm{E}[Y \mid r_{1:N}, \alpha, \beta_{1:K}, \pi_{1:K}, \eta]$.

Variational approximate inference

Both Parameter estimation and prediction depend on the posterior inference. Following Jordan et al. [16], given patient HDRs and response variable y, we start the joint distribution in the following equation. The real posterior of latent variables (including topic proportions θ, topic r_n) is:

$$p\left(\theta, r_{1:N} \mid f_{1:N}, \nu_{1:N}, y, \alpha, \pi_{1:\mathrm{K}}, \beta_{1:K}, \eta\right) = \frac{p(\theta \mid \alpha)\left(\prod_{n=1}^{N} p(\mathrm{r}_n \mid \theta) p\left(f_n \mid \mathrm{r}_n, \pi_{1:\mathrm{K}}\right) p\left(\nu_n \mid \mathrm{r}_n, f_n, \beta_{1:K}\right)\right) p\left(y \mid r_{1:N}, \eta\right)}{\int p(\theta \mid \alpha) d\theta \sum_{r_{1:\mathrm{N}}}\left(\prod_{n=1}^{N} p(\mathrm{r}_n \mid \theta) p\left(f_n \mid \mathrm{r}_n, \pi_{1:\mathrm{K}}\right) p\left(\nu_n \mid \mathrm{r}_n, f_n, \beta_{1:K}\right)\right) p\left(y \mid r_{1:N}, \eta\right)}$$

As the denominator of posterior distribution is difficult to calculate, we use mean-field variational approximation inference:

$$q(\theta, r_{1:N} \mid \gamma, \phi_{1:\mathrm{N}}) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(\mathrm{r}_n \mid \phi_n)$$

Let $\zeta = \{\alpha, \pi_{1:K}, \beta_{1:K}, \eta\}$, the *KL* divergence between the real posterior of latent variables $p\left(\theta, r_{1:N} \mid f_{1:N}, \nu_{1:N}, y, \zeta\right)$ and variational distribution $q(\theta, r_{1:\mathrm{N}})$ is:

$$\begin{aligned}
&\mathrm{D}\left(q(\theta, r_{1:N}) \| p\left(\theta, r_{1:N} \mid f_{1:N}, \nu_{1:N}, y, \zeta\right)\right) \\
&\quad = \mathrm{E}_q[\log q(\theta, r_{1:N})] - \mathrm{E}_q\left[\log p\left(\theta, r_{1:N} \mid f_{1:N}, \nu_{1:N}, y, \zeta\right)\right] \\
&\quad = \mathrm{E}_q[\log q(\theta, r_{1:N})] - \mathrm{E}_q\left[\log p\left(\theta, r_{1:N}, f_{1:N}, \nu_{1:N}, y \mid \zeta\right)\right] \\
&\qquad + \log p\left(f_{1:N}, \nu_{1:N}, y \mid \zeta\right) \\
&\quad \geq 0
\end{aligned}$$

So the evidence lower bound (*ELBO*) is:

$$\begin{aligned}
&\log p\left(f_{1:N}, \nu_{1:N}, y \mid \zeta\right) \\
&\geq \mathrm{E}_q\left[\log p\left(\theta, r_{1:N}, f_{1:N}, \nu_{1:N}, y \mid \zeta\right)\right] \\
&\quad - \mathrm{E}_q[\log q(\theta, r_{1:N})]
\end{aligned}$$

we denote *ELBO* by $\mathscr{L}(\bullet)$, and the entropy of variational distribution by $H(q) = -\mathrm{E}_q[\log q(\theta, r_{1:N})]$, and then *ELBO* is written as:

$$\begin{aligned}
&\mathscr{L}\left(f_{1:N}, \nu_{1:N}, y \mid \zeta\right) \\
&\quad = \mathscr{L}(\gamma, \phi_{1:N} \mid \zeta) = \mathrm{E}_q[\log p(\theta \mid \alpha)] \\
&\qquad + \sum_{n=1}^{N} \mathrm{E}_q[\log p(\mathrm{r}_n \mid \theta)] \\
&\qquad + \sum_{n=1}^{N} \mathrm{E}_q\left[\log p\left(f_n \mid \mathrm{r}_n, \pi_{1:\mathrm{K}}\right)\right] \\
&\qquad + \sum_{n=1}^{N} \mathrm{E}_q\left[\log p\left(\nu_n \mid \mathrm{r}_n, f_n, \beta_{1:\mathrm{F}}\right)\right] \\
&\qquad + \mathrm{E}_q\left[\log p\left(y \mid r_{1:N}, \eta\right)\right] + H(q)
\end{aligned} \tag{1}$$

We fit these parameters by maximizing *ELBO* with respect to γ, ϕ and obtain an estimate of the posterior under the sense of *KL* divergence between $q(\theta, r_{1:N})$ and the true posterior $p\left(\theta, r_{1:N} \mid f_{1:N}, v_{1:N}, y, \zeta\right)$.

The terms of equation 1 are as follows:

$$\mathrm{E}_q[\log p(\theta \mid \alpha)] = \log \Gamma\left(\sum_{i=1}^{K} \alpha_i\right) - \sum_{i=1}^{K} \log \Gamma(\alpha_i) + \sum_{i=1}^{K} (\alpha_i - 1) E_q[\log \theta_i] \tag{2}$$

$$\sum_{n=1}^{N} \mathrm{E}_q[\log p(\mathrm{r}_n \mid \theta)] = \sum_{\mathrm{n}=1}^{N} \sum_{i=1}^{K} \varphi_{n,i} \mathrm{E}_{q(\theta_i \mid \gamma)}[\log \theta_i] \tag{3}$$

$$\sum_{n=1}^{N} \mathrm{E}_q\left[\log p\left(f_n \mid \mathrm{r}_n, \pi_{1:\mathrm{K}}\right)\right] = \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_{n,i} \log \pi_{i,f_n} \tag{4}$$

$$\sum_{n=1}^{N} \mathrm{E}_q\left[\log p\left(v_n \mid \mathrm{r}_n, f_n, \beta_{1:\mathrm{k}}\right)\right] = \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_{n,i} \log \beta_{i,v_n} \tag{5}$$

$$E_q[\log p\left(y \mid r_{1:N}, \eta\right)] = \eta_y^T \frac{1}{N} \sum_{n=1}^{N} \phi_n - E_q\left[\log \sum_{l=1}^{C} \exp\left(\eta_l^T \bar{\mathrm{r}}\right)\right] \tag{6}$$

$$\begin{aligned} H(q) = &- \log \Gamma\left(\sum_{i=1}^{K} \gamma_i\right) + \sum_{i=1}^{K} \log \Gamma(\gamma_i) \\ &- \sum_{i=1}^{K} (\gamma_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\sum_{l=1}^{K} \gamma_l\right)\right) \\ &- \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_{n,i} \log \phi_{n,i} \end{aligned} \tag{7}$$

where $\Psi(\bullet)$ denotes the digamma function.

Variational E-step

The coordinate ascent method updates for variational parameter γ, which is the same as sLDA, does not directly involve the response variable y.

$$\gamma^{new} \leftarrow \alpha + \sum_{n=1}^{N} \phi_n \tag{8}$$

Under the conditions $E[\bar{\mathrm{r}}] = \bar{\phi} = \frac{1}{N} \sum_{n=1}^{N} \phi_n$, the terms in *ELBO* containing ϕ_n are:

$$\begin{aligned} \mathscr{L}_{[\phi_n]} = &\sum_{i=1}^{K} \phi_{n,i} E_q[\log \theta_i] + \sum_{i=1}^{K} \phi_{n,i} \log \pi_{i,f_n} \\ &+ \sum_{i=1}^{K} \phi_{n,i} \log \beta_{i,v_n} - \sum_{i=1}^{K} \phi_{n,i} \log \phi_{n,i} \\ &+ \frac{1}{N} \sum_{i=1}^{K} \eta_{y,i} \phi_{n,i} \\ &- \log\left(\sum_{l=1}^{C} \prod_{n=1}^{N}\left(\sum_{i=1}^{K} \phi_{n,i} \exp\left(\eta_{l,i} \frac{1}{N}\right)\right)\right) \end{aligned}$$

Following Wang et al. [22], under the constraint $\sum_{i=1}^{K} \phi_{n,i} = 1$ and setting partial derivatives to zero of the *ELBO* with respect to $\phi_{n,i}$, we write $\log\left(\sum_{l=1}^{C} \prod_{n=1}^{N}\left(\sum_{i=1}^{K} \phi_{n,i} \exp\left(\eta_{l,i} \frac{1}{N}\right)\right)\right)$ as $h^T \phi_n$, and then obtain:

$$\phi_{n,i} \propto \pi_{i,f_n} \beta_{i,v_n} \exp\left(\Psi'(\gamma_i) + \frac{\eta_{y,i}}{N} - \left(h^T \phi_n^{old}\right)^{-1} h_i\right) \tag{9}$$

The variational EM algorithm alternatively updates Eqs. 8 and 9 until the bound on the expected log likelihood converges.

Parameter estimation

Variational M-step is an optimization of *ELBO* on the whole HDRs datasets level w.r.t model parameters $\zeta = \{\alpha, \pi_{1:K}, \beta_{1:K}, \eta\}$. Repeating the Variational E-step D times, we can obtain approximate posterior over latent variables $\theta, r_{1:N}$ for each patient HDRs. It is noted that different patient HDRs has different variational distributions $q_d(\theta, r_{1:N})$. Then we obtain:

$$\begin{aligned} &L(\alpha, \pi_{1:\mathrm{K}}, \beta_{1:K}, \eta; D) \\ &= \sum_{\mathrm{d}=1}^{D} \{E_{q_d}\left[\log p\left(\theta_d, \mathrm{r}_{d,1:N}, f_{d,1:N}, v_{d,1:N}, y_d\right)\right] \\ &\quad + H(q_d)\} \end{aligned}$$

Variational M-step

Similarly, the coordinate ascent method is used to maximize the whole HDRs datasets *ELBO* to estimate the model parameters $\zeta = \{\alpha, \pi_{1:K}, \beta_{1:K}, \eta\}$.

(a) Setting

$$\partial L(\alpha, \pi_{1:\mathrm{K}}, \beta_{1:\mathrm{K}}, \eta; D)/\partial \pi_{k,f} = 0$$
$$\partial L(\alpha, \pi_{1:\mathrm{K}}, \beta_{1:\mathrm{K}}, \eta; D)/\partial \beta_{k,\mathrm{v}} = 0$$

it leads to:

$$\hat{\pi}_{k,f}^{new} \propto \sum_{d=1}^{D} \sum_{n=1}^{N} I\left(f = f_{n,k}^{d}\right) \phi_{n,k}^{d} \tag{10}$$

$$\hat{\beta}_{k,v}^{new} \propto \sum_{d=1}^{D} \sum_{n=1}^{N} I\left(v = v_{n,k}^{d}\right) \phi_{n,k}^{d} \tag{11}$$

(b) The whole HDRs datasets *ELBO* containing η_c are:

$$L_{[\eta_{1:C}]}(\mathrm{D}) = \sum_{d=1}^{D} \left(\eta_{c_d}^{T} \bar{\phi}_d - \log \left(\sum_{l=1}^{C} \left\{ \prod_{n=1}^{N} \left(\sum_{i=1}^{K} \phi_{n,i}^{d} \exp \left(\eta_{l,i} \frac{1}{N} \right) \right) \right\} \right) \right)$$

Setting $\frac{\partial L_{[\eta_{1:C}]}(\mathrm{D})}{\partial \eta_{\mathrm{c},i}} = 0$ does not lead to a closed-form solution. Following Wang et al. [22], we optimize with conjugate gradient. Let $\kappa_d = \sum_{l=1}^{C} \left\{ \prod_{n=1}^{N} \phi_{n,\mathrm{r}_n} \sum_{\mathrm{i}=1}^{K} \left(\exp \left(\eta_{l,i} \frac{1}{N} \right) \right) \right\}$, the derivatives are:

$$\begin{aligned} \frac{\partial L_{[\eta_{1:\mathrm{c}}]}(\mathrm{D})}{\partial \eta_{\mathrm{c},i}} &= \sum_{d=1}^{D} \left(1[\mathrm{c}_d = \mathrm{c}] \bar{\phi}_{d,i} \right) \\ &- \sum_{d=1}^{D} \left(\kappa_d^{-1} \prod_{n=1}^{N} \left(\sum_{j=1}^{K} \phi_{n,\mathrm{j}}^{\mathrm{d}} \exp \left(\eta_{\mathrm{c}_j} \frac{1}{N} \right) \right) \right. \\ &\times \left. \sum_{n=1}^{N} \left(\frac{\frac{1}{N} \phi_{n,\mathrm{i}}^{\mathrm{d}} \exp \left(\eta_{\mathrm{c}_i} \frac{1}{N} \right)}{\sum_{j=1}^{K} \phi_{n,\mathrm{j}}^{\mathrm{d}} \exp \left(\eta_{\mathrm{c}_j} \frac{1}{N} \right)} \right) \right) \end{aligned}$$

(c) In this paper, We will address this setting in details about the Dirichlet parameter α and β in "Discussion" section.

Prediction and risk stratification

Under the fitted model$\{\alpha, \gamma, \phi_{1:N}, \pi_{1:K}, \beta_{1:K}, \eta\}$, the expected response value is:

$$\begin{aligned} &E[Y \mid f_{1:N}, v_{1:N}, \alpha, \pi_{1:K}, \beta_{1:K}, \eta] \\ &= E\left[\mu\left(\eta^{\top} \bar{r}\right) \mid f_{1:N}, v_{1:N}, \alpha, \pi_{1:K}, \beta_{1:K}\right] \end{aligned}$$

where $\mu(\bullet) = \mathrm{E}_{GLM}[Y \mid \cdot] = \left[\frac{\exp\left(\eta_1^T \bar{\mathrm{r}}\right)}{\sum_{l=1}^{C} \exp\left(\eta_l^T \bar{\mathrm{r}}\right)}, \ldots, \frac{\exp\left(\eta_C^T \bar{\mathrm{r}}\right)}{\sum_{l=1}^{C} \exp\left(\eta_l^T \mathrm{r}\right)} \right]^T$.

In classification, to estimate the probability of the label c with the variational distribution, we obtain:

$$\begin{aligned} &E[Y \mid v_{1:N}, \alpha, \pi_{1:K}, \ \beta_{1:K}, \eta] \\ &\approx \int \exp \left(\log \frac{\exp\left(\eta_c^T \bar{\mathrm{r}}\right)}{\sum_{l=1}^{C} \exp\left(\eta_l^T \mathrm{r}\right)} q(r) \right) \mathrm{dr} \\ &\geq \exp \left(\mathrm{E}_q \left[\eta_c^T \bar{r} \right] - \mathrm{E}_q \left[\log \left(\sum_{l=1}^{C} \exp \left(\eta_l^T \bar{r} \right) \right) \right] \right) \end{aligned}$$

Thus, the prediction formulation is:

$$\mathrm{c}^* = \arg \max_{c \in \{1,\ldots,\mathrm{C}\}} \mathrm{E}_q \left[\eta_c^T \bar{r} \right] = \arg \max_{c \in \{1,\ldots,\mathrm{C}\}} \eta_c^T \bar{\phi} \tag{12}$$

where $E[\bar{r}] = \bar{\phi} = \frac{1}{N} \sum_{n=1}^{N} \phi_n$.

Experiments

Methods

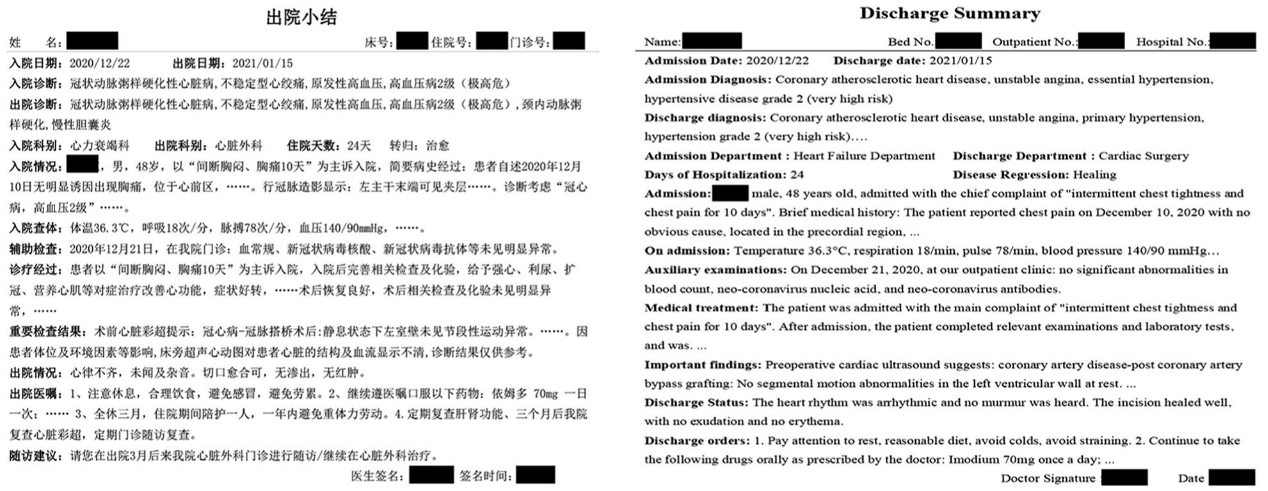
Data source

We propose our model on real-world datasets in the clinical domains collected from the Cardiology Department of the First Affiliated Hospital of Xinjiang Medical University containing 420 HDRs of CHD patients. These HDRs describe basic information about patients with a first diagnosis of coronary atherosclerotic heart disease, such as admission, treatment history, important test results, discharge status, discharge orders, and follow-up recommendations. An original HDR of a patient was described in Fig. 2-left, corresponding English version was shown in Fig. 2-right. In this paper, the patient's diagnostic results is used as the class label for the data, including the following 4 classes: Stable Angina Pectoris(SAP), Unstable Angina Pectoris(UAP), Ischemic Cardiomyopathy(ICM) and Acute Miocardial Infarction(AMI).

Annotation of features

HDRs are a kind of unstructured text. In order to guarantee the accuracy of datasets and ensure the credibility as far as possible, each patient case in the experimental datasets was identified and evaluated by clinicians. This is a labor-intensive and laborious task.

With an annotation tool developed by ourselves, we finished the features annotation of patients and obtain structured data including patient demographics, general items, history of present illness (HPI), laboratory results, conclusions of diagnoses etc., which provide a comprehensive source for risk stratification. Based on the CHD risk factors formulated in I2B2 and combined with the clinical experience of clinicians in the diagnosis and treatment, we drew up annotation criterion and guidelines for this study under the help of clinicians.



出院小结

姓　　名：[redacted]　床号：[redacted] 住院号：[redacted] 门诊号：[redacted]

入院日期： 2020/12/22　**出院日期：** 2021/01/15

入院诊断： 冠状动脉粥样硬化性心脏病，不稳定型心绞痛，原发性高血压，高血压病2级（极高危）

出院诊断： 冠状动脉粥样硬化性心脏病，不稳定型心绞痛，原发性高血压，高血压病2级（极高危），颈内动脉粥样硬化，慢性胆囊炎

入院科别： 心力衰竭科　**出院科别：** 心脏外科　**住院天数：** 24天　转归：治愈

入院情况： [redacted]，男，48岁，以“间断胸闷、胸痛10天”为主诉入院，简要病史经过：患者自述2020年12月10日无明显诱因出现胸痛，位于心前区，……。行冠脉造影显示：左主干末端可见夹层……。诊断考虑“冠心病，高血压2级”……。

入院查体： 体温36.3℃，呼吸18次/分，脉搏78次/分，血压140/90mmHg，……。

辅助检查： 2020年12月21日，在我院门诊：血常规、新冠状病毒核酸、新冠状病毒抗体等未见明显异常。

诊疗经过： 患者以“间断胸闷、胸痛10天”为主诉入院，入院后完善相关检查及化验，给予强心、利尿、扩冠、营养心肌等对症治疗改善心功能，症状好转，……术后恢复良好，术后相关检查及化验未见明显异常，……

重要检查结果： 术前心脏彩超提示：冠心病-冠脉搭桥术后：静息状态下左室壁未见节段性运动异常。……。因患者体位及环境因素等影响，床旁超声心动图对患者心脏的结构及血流显示不清，诊断结果仅供参考。

出院情况： 心律不齐，未闻及杂音。切口愈合可，无渗出，无红肿。

出院医嘱： 1、注意休息，合理饮食，避免感冒，避免劳累。2、继续遵医嘱口服以下药物：依姆多 70mg 一日一次；…… 3、全休三月，住院期间陪护一人，一年内避免重体力劳动。4. 定期复查肝肾功能、三个月后我院复查心脏彩超，定期门诊随访复查。

随访建议： 请您在出院3月后来我院心脏外科门诊进行随访/继续在心脏外科治疗。

医生签名：[redacted] 签名时间：[redacted]

Discharge Summary

Name:[redacted]　Bed No.[redacted] Outpatient No.:[redacted] Hospital No.:[redacted]

Admission Date: 2020/12/22　**Discharge date:** 2021/01/15

Admission Diagnosis: Coronary atherosclerotic heart disease, unstable angina, essential hypertension, hypertensive disease grade 2 (very high risk)

Discharge diagnosis: Coronary atherosclerotic heart disease, unstable angina, primary hypertension, hypertension grade 2 (very high risk)….

Admission Department : Heart Failure Department　**Discharge Department :** Cardiac Surgery

Days of Hospitalization: 24　**Disease Regression:** Healing

Admission:[redacted] male, 48 years old, admitted with the chief complaint of "intermittent chest tightness and chest pain for 10 days". Brief medical history: The patient reported chest pain on December 10, 2020 with no obvious cause, located in the precordial region, ...

On admission: Temperature 36.3°C, respiration 18/min, pulse 78/min, blood pressure 140/90 mmHg…

Auxiliary examinations: On December 21, 2020, at our outpatient clinic: no significant abnormalities in blood count, neo-coronavirus nucleic acid, and neo-coronavirus antibodies.

Medical treatment: The patient was admitted with the main complaint of "intermittent chest tightness and chest pain for 10 days". After admission, the patient completed relevant examinations and laboratory tests, and was. ...

Important findings: Preoperative cardiac ultrasound suggests: coronary artery disease-post coronary artery bypass grafting: No segmental motion abnormalities in the left ventricular wall at rest. ...

Discharge Status: The heart rhythm was arrhythmic and no murmur was heard. The incision healed well, with no exudation and no erythema.

Discharge orders: 1. Pay attention to rest, reasonable diet, avoid colds, avoid straining. 2. Continue to take the following drugs orally as prescribed by the doctor: Imodium 70mg once a day; ...

Doctor Signature [redacted]　Date [redacted]

Fig. 2 Original HDR. The original HDR in Chinese (left); The corresponding English version (right)

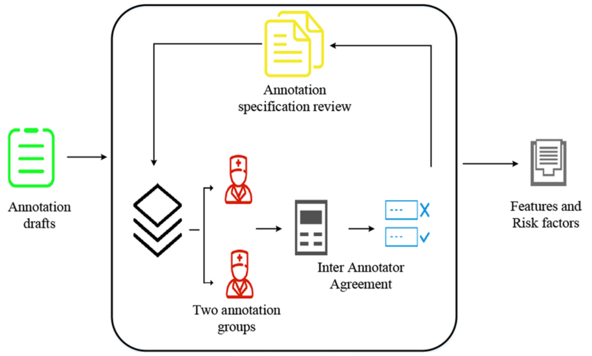

Annotation specification review
Annotation drafts
Features and Risk factors
Inter Annotator Agreement
Two annotation groups

Fig. 3 Features annotation of HDRs. The process of features annotation of HDRs

This work consists of three pre-annotation and a formal annotation, with 50 HDRs randomly selected each time. The result was verified by Inter Annotator Agreement (IAA) after each annotation to guarantee the qualification. At the same time, the criterion and guidelines of annotation were constantly updated to ensure the standardization throughout the process. Figure 3 shows the features annotation process of HDRs.

Data preprocessing

While random forests (RFs) [30], which allow for avoiding over-fitting makes it suitable for processing data with outliers, missing values, have been widely applied to other fields such as biological prediction [31], we use RFs for filling the missing data.

And then, we obtain 34 patient features and 79 patient feature-value pairs, according to specific medical guidelines and specifications of CHD, by dividing patient features into a series of categories. Using the TF-IDF text mining technique to assign a weight to each feature and feature-value pair term, features matrix of 420 × 34 dimensions and feature-value pairs matrix of 420 × 79 dimensions are obtained respectively as the source datasets for modeling.

At the same time, 265 out of the 420 patient HDRs cases in the datasets are Stable Angina Pectoris (63.10%), and the class imbalance problem is encountered. In this paper, we use the SMOTE algorithm to generate additional samples from randomly oversampling minority class, and finally totaling 1060 samples are obtained. Therefore, features and diagnosis results can be extracted from HDRs, which provide data preparation for supervised training and testing and can be used as the labels of CHD. Summary statistics of the datasets are shown in Table 2.

Results

For risk stratification and classification prediction of CHD, we use fivefold cross-validation method to create the train and test sets. Results are reported as an average across folds.

Classification performance

In order to perform assessment of the classification performance of our model, we compared it with previous salient approach Multi-class sLDA [22]. Multi-class sLDA embeds single softmax into LDA model, and reports better classification performance. The distinguishing factor between Hierarchical sLDA and Multi-class sLDA is the additional structure imposed on the feature-value pair, which would result in an outstanding performance in predictive performance.

Table 2 Summary statistics of datasets

Number of patient records		**Number of patient features**		**Number of patient feature-value pairs**
420		34		79
Top 15 risk factors		**Frequency**	**Ratio (%)**	**Descriptions**
Heart rhythm-Sinus		404	96.19%	
Antiplatelet medication-Yes		393	93.57	
Heart rate-norm.		371	88.33	
Lipid-lowering medication-Yes		369	87.86	
Chest pain-1*		338	80.48	1. Oppressive, stuffy or constrictive* 2. Dyspnea
Range of Chest pain-2*		306	72.86	1. Located behind the sternal body 2. Affected precordial area, palm size range* 3. Radiation to the left shoulder, left arm medial ring finger and little finger
Sex-Male		288	68.57	
Cardiac B-Ultrasound-Abn.		279	66.43	
Hypertension-Yes		271	64.52	
Ethnic-Han		253	60.24	
Incentive-Yes		224	53.33	
LDL-C-Abn.		253	60.24	
PCI or CABG-Yes		223	53.10	PCI: percutaneous coronary intervention CABG: coronary artery bypass grafting
β blocker medications-Yes		170	40.48	
Carotid atherosclerosis with plaque-2*		164	39.05	1. No atherosclerotic plaque 2. Single plaque group* 3. Multiple plaque group
Types of CHD	SAP	265	63.10	
	UAP	98	23.33	
	ICM	12	2.86	
	AMI	45	10.71	

The experiments were performed from topic $K = 10$ to $K = 70$ with intervals of 10. The results of accuracy, training and testing time are illustrated in Fig. 4 and the confusion matrices are shown in Fig. 5. Hierarchical sLDA model we proposed is competitive in both time performance and accuracy, as validated by experimental results. From Fig. 4-left and middle, comparison of over all classes based on fivefold cross validation it can be seen that Hierarchical sLDA ($K = 70$) reaches the average training time (669.69s) and testing time (0.32s) performance, which is 59.74% and 25.58% higher than the average training time (1663.42s) and testing time (0.43s) of Multi-class sLDA ($K = 70$) but almost no loss of accuracy on our datasets (See Fig. 4-right). The average accuracy of the Hierarchical sLDA ($K = 70$) is 74.53%, while that of Multi-class sLDA ($K = 70$) is 74.06%.

From Fig. 4-right, the average classification accuracy of the Hierarchical sLDA, as the number of topics increases, is smoothly converging, as well as not suffering from the overfitting problem. That means it provides more robust than other classifiers.

We evaluated on topic $K = 70$ about different types of CHD using macro-F1 score, macro-Precision score, and macro-Recall score on test data respectively. A comparison between Hierarchical sLDA model and Multi-class sLDA model indicated that the two models are not significantly different(See Table 3).

Risk stratification

Table 4 shows the top 5 risk factors of $K = 70$ topics inferred by Hierarchical sLDA under high-, and low-risk tier separately of different types of CHD. The Hierarchical sLDA model of HDRs for CHD shows us:

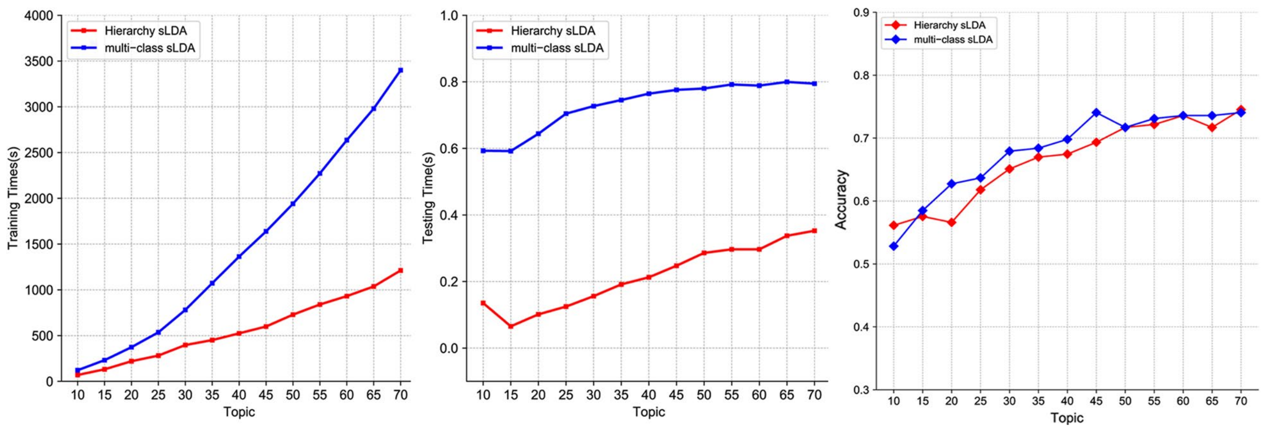


Fig. 4 Comparison of performance. Comparison of over all classes based on fivefold cross validation: training time (left); testing time (middle); average accuracy (right)

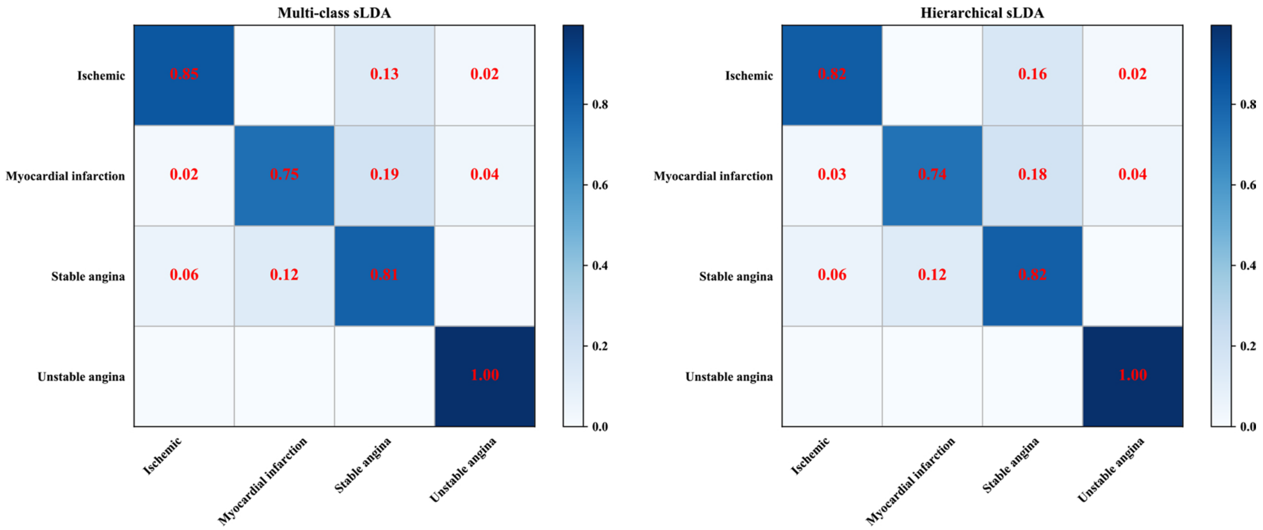


Fig. 5 Comparison of confusion matrices. Comparison of confusion matrices of topic $K = 70$; multi-class sLDA (left); Hierarchical sLDA (right)

Table 3 Comparison of macro-F1, Precision, Recall

	Macro-F1 (%)	Macro-Precision (%)	Macro-Recall (%)
Hierarchical sLDA	70.91	70.96	71.70
Multi-class sLDA	70.54	70.84	71.70

(a) Different types of CHD generally have different risk factors, while Diabetes-Yes and Antiplatelet Medication-Yes are risk factors of Stable angina pectoris and Unstable angina pectoris separately. We can take a conclusion that Antiplatelet Medication provides an effective treatment, simultaneously should be highly aware of Diabetes.

(b) Uric acid-Abn. and SBP-Abn. are the risk factors of 4 types of CHD, which indicate the cause-and-effect correlation between these two risk factors and CHD.

(c) Both antiplatelet medication, ACEI/ARB and Lipid drug therapy are often used to reduce high-risk factors.

(d) Gender seems to have higher probability of different types of CHD.

Table 4 Risk factors of CHD extracted from Hierarchical sLDA

ICM	AMI	SAP	UAP
High-risk			
ST segment-Abn.	Hb-Abn.	SBP-Abn.	Diabetes-Yes
CTA stenosis-Mild	Duration-10 min	HbA1c-Abn.	Antiplatelet Medication-Yes
Carotid Atherosclerosis-Multiple plaque group	ST segment-Elevation	Uric acid-Abn.	HbA1c-Abn.
Fasting blood glucose-Abn.	β blocker medications-Yes	ST segment-Change	Duration-3~5 min
Heart rate-Sinus velocity	Gender-Female	Lipid drug medication-Yes	Gender-Male
Low-risk			
CTA lesions-Single	CTA lesions-Single	Cardiac B-Ultrasound -Abn.	Hypertension-Yes
HbA1c-Abn.	ACEI/ARB medication-Yes	Hb-Abn.	DBP-Abn.
Age-45–65 years	Age-45–65 years	CTA stenosis-Mild	WBC-Abn.
Uric acid-Abn.	Uric acid-Abn.	Diabetes-Yes	SBP- Abn.
SBP- Abn.	SBP-Abn.	Antiplatelet medication-Yes	Uric acid-Abn.

Dirichlet hyperparameter

According to experience, Wei and Croft [32] pointed out that the choice of Dirichlet hyperparameter is not sensitive to the experimental results. They used symmetric Dirichlet priors in the estimation $\alpha = 50/K$ and $\beta = 0.01$. In this paper, according to our analysis (See "Parameter estimation" section) and the experimental results in Fig. 6, under the same topic (e.g., $K = 70$), selecting different hyperparameter α has no significant effect on the average prediction accuracy (Fig. 6-right). However, after using α from 0.1 to 1.5 with interval of 0.1, K from 10 to 70 with interval of 5, the 3-D represention from Fig. 6-left and -right show that, rather than $\alpha = 50/K$, we should optimize hyperparameter α with the fitted curve

$$\alpha = c_1 + \frac{1}{1 + exp\left(c_2 * \left(K - \frac{|V|}{2}\right)\right)}$$

where c_1 and c_2 are constants and $|V|$ is the number of patient feature-value pairs (e.g., $c_1 = 0.3$, $c_2 = 0.25$ and $|V| = 79$ in our experiments).

Panichella [33] also showed that search-based approaches are very effective in solving LDA hyperparameter tuning problem. In our proposed Hierarchical sLDA model, Dirichlet hyperparameter can affect the speed of convergence (See "Parameter estimation"

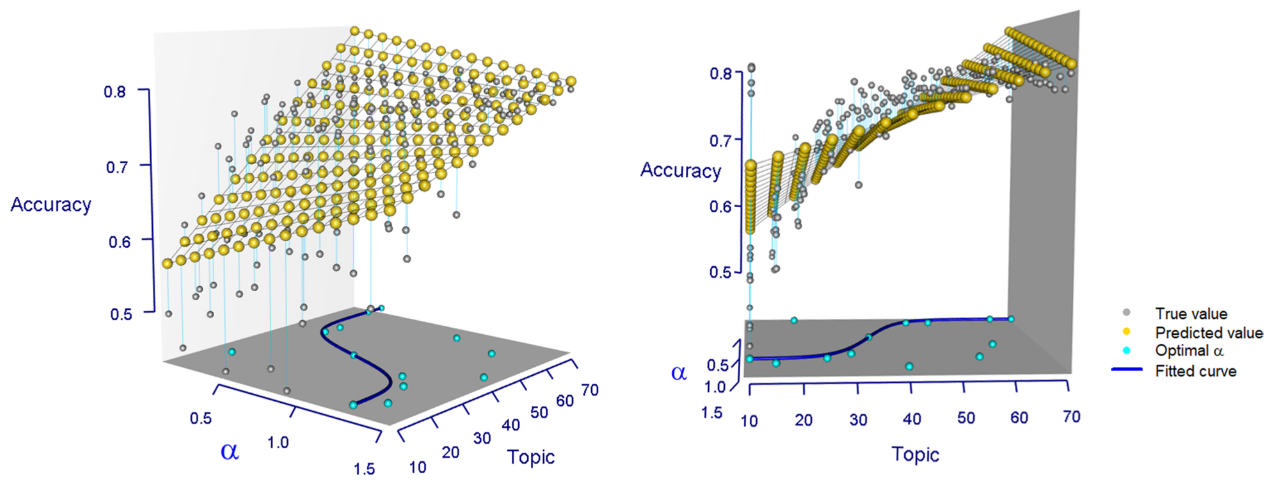


Fig. 6 Selection of optimal hyperparameter α. Using α from 0.1 to 1.5 with interval of 0.1, topic K from 10 to 70 with interval of 5, the 3-D represention show that we should optimize hyperparameter α with the fitted curve. The left view of 3-D represention (left); the front view of 3-D represention (right)

section, Eqs. 10 and 11), therefore this structure has the potential to become a generic scheme for variants of sLDA-based model.

Discussion

Traditional statistical models for risk stratification of CVD risk are well-developed, but typically less flexible than machine learning techniques and only hold under well controlled conditions for prediction and classification.

Hierarchical sLDA model is a variant of sLDA. The experimental results clearly demonstrate that the proposed Hierarchical sLDA has significant advantages on supervised CHD classification and risk stratification relative to the compared Multi-class sLDA approach, including accuracy and time performance.

Most classification techniques do not handle hierarchical features, which offer little insight beyond a flat feature-based segmentation, as they assume that features in the training datasets are fully independent. By categorizing patient features in HDRs as patient feature-value pairs, three-layer sLDA is improved to four-layer Hierarchical sLDA, which can accelerate the convergence of time-consuming Gibbs sampling.

We have shown that the model, as the number of topics increases, is converging smoothly. That means it is not suffering from overfitting problem and provides more robust than other classifiers.

Intriguingly, according to experience, Dirichlet hyperparameters are set to $\alpha = 50/K$ and $\beta = 0.01$. However, we recommend that Dirichlet hyperparameter α may be optimized in another setup policy (See "Dirichlet hyperparameter" section). Through experiment analysis, there exists two limitations and weaknesses for the current approach:

(a) Insufficient training data. From the data source process, due to specific difficulties, features extraction is semi-automatic, which can be time and labor intensive. And therefore, Insufficient training data limits the performance of the model.
(b) More experiments. For future work, we will investigate the performance of our model when applied to other topic models and datasets.

Conclusions

We hereby have proposed an approach in HDRs for risk stratification and classification of CHD simultaneously over our datasets, which is competitive in time performance and accuracy. Hierarchical processing of patient features can significantly improve the disadvantages of low efficiency and time-consuming Gibbs sampling of sLDA model. Meanwhile our model has the potential to be applied to other datasets by transforming the features of datasets into feature-value pairs. On the other hand, while coronary angiography is the gold standard for the diagnosis of CHD, but its limitations, such as invasive, random errors by the selection of radiographic projection, limit its wide clinical applications. Risk factors, which is extracted from risk stratification, can be used as a reference to provide individualized prevention and therapeutic decisions with non-invasive methods. However, the difficulty in processing complicated clinical applications suggests that this is still an open question needed to be solved in future research.

Acknowledgements
The authors sincerely thank the cooperative hospital as well as all medical staffs involved for their supports. In addition, we would like to thank the anonymous reviewers for comments and suggestions offered for improvements.

Authors' contributions

GY wrote the main manuscript text. YZ prepared datasets. LZ, JZ, TZ designed the experiment. XB prepared data preprocessing. All authors read and approved the final manuscript.

Funding
This work was supported by the Natural Science Foundation of Xinjiang Uygur Autonomous Region under Grant No. 2019D01C205 and No. 2019D01C041. This content is entirely the responsibility of the authors and does not represent the official views of founder.

Availability of data and materials

The datasets used in this study and the source code of Hierarchical sLDA are available at https://github.com/flyingbiao/Hierachical-sLDA. All experiments were performed on Intel(R) Core(TM) i7 @4.20GHz with 16G bytes main memory running on Ubuntu 18.04.2. The algorithm was implemented using C++.

Declarations

Ethics approval and consent to participate

This study was approved by the ethics committee of the Medical Ethics Committee of the First Affiliated Hospital of Xinjiang Medical University, and the committee's reference number is K202107-01.

Consent for publication

Not applicable.

Competing interests

The authors report that they have no conflicts.

Author details
[1]School of Medical Engineering and Technology, Xinjiang Medical University, No.567 North Shangde Road, Urumqi, China. [2]College of Information Science and Engineering, Xinjiang University, Urumqi, China. [3]The First Affiliated Hospital of Xinjiang Medical University, Urumqi, China.

Received: 22 June 2021 Accepted: 5 January 2022
Published online: 15 January 2022

References

1. Rod J, Carlene Mm L, et al. Treatment with drugs to lower blood pressure and blood cholesterol based on an individual's absolute cardiovascular risk. Lancet. 2014;384(9943):591–8. https://doi.org/10.1016/S0140-6736(14)61212-5.
2. Schlesinger DE, Stultz CM. Deep learning for cardiovascular risk stratification. Curr Treat Options Cardiovasc Med. 2020. https://doi.org/10.1007/s11936-020-00814-0.
3. Brindle P, Beswick A, Fahey T, Ebrahim S. Accuracy and impact of risk assessment in the primary prevention of cardiovascular disease: a systematic review. Heart. 2006;92(12):1752–9. https://doi.org/10.1136/hrt.2006.087932.
4. Matheny M, Mcpheeters ML, et al. Systematic review of cardiovascular disease risk assessment tools [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2011.
5. Hsueh PYS, Zhu XX, et al. Automatic summarization of risk factors preceding disease progression an insight-driven healthcare service case study on using medical records of diabetic patients. World Wide Web Internet Web Inf Syst. 2015;18(4):1163–75. https://doi.org/10.1007/s11280-014-0304-2.
6. Brom H, Brooks Carthon JM, Ikeaba U, Chittams J. Leveraging electronic health records and machine learning to tailor nursing care for patients at high risk for readmissions. J Nurs Care Qual. 2019;35(1):27–33. https://doi.org/10.1097/NCQ.0000000000000412.
7. Whitlock EL, Braehler MR, Kaplan JA, Finlayson E, Rogers SE, Douglas V, Donovan AL. Derivation, validation, sustained performance, and clinical impact of an electronic medical record-based perioperative delirium risk stratification tool. Anesth Analg. 2020;131(6):1901–10. https://doi.org/10.1213/ANE.0000000000005085.
8. Safarova MS, Kullo IJ. Using the electronic health record for genomics research. Curr Opin Lipidol. 2020;31(2):85–93. https://doi.org/10.1097/MOL.0000000000000662.
9. Petersen JD, Lozovatsky M, Markovic D, Duncan R, Zheng S, Shamsian A, Kagele S, Ross MK. Clinical decision support for hyperbilirubinemia risk assessment in the electronic health record. Acad Pediatr. 2020;20(6):857–62. https://doi.org/10.1016/j.acap.2020.02.009.
10. Stubbs A, Kotfila C, Xu H, Uzuner O. Identifying risk factors for heart disease over time: overview of 2014 i2b2/UTHealth shared task Track 2—ScienceDirect. J Biomed Inform. 2015;58(S):67–77. https://doi.org/10.1016/j.jbi.2015.07.001.
11. Conroy R, Sans S, Fitzgerald A, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the score project. Eur Heart J. 2003;11(24):987–1003. https://doi.org/10.1016/s0195-668x(03)00114-3.
12. Goff DC, Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. J Am Coll Cardiol. 2014;63(25,B):2935–59. https://doi.org/10.1016/j.jacc.2013.11.005.
13. Huang Z, Dong W, Duan H. A probabilistic topic model for clinical risk stratification from electronic health records. J Biomed Inform. 2015;58:28–36. https://doi.org/10.1016/j.jbi.2015.09.005.
14. Blei DM. Probabilistic topic models. Commun ACM. 2012. https://doi.org/10.1145/2133806.2133826.
15. Hofmann T. Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval—SIGIR '99 0. 1999. p. 50–57. https://doi.org/10.1145/312624.312649.
16. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res. 2003;3:993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993.
17. Girolami M, Kabán A. On an equivalence between PLSI and LDA. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval—SIGIR '03. 2003. p. 433–434. https://doi.org/10.1145/860435.860537.
18. Steyvers M, Griffiths T. Handbook of latent semantic analysis. 2014. p. 427–448. https://doi.org/10.4324/9780203936399.ch21.
19. Jelodar H, Wang Y, Yuan C, Feng X. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed Tools Appl. 2017;78:15169–211. https://doi.org/10.1007/s11042-018-6894-4.
20. Li W, McCallum A. Pachinko allocation: DAG-structured mixture models of topic correlations. In: Proceedings of the 23rd international conference on machine learning—ICML'06. 2006. p. 577–584. https://doi.org/10.1145/1143844.1143917.
21. Blei DM, McAuliffe JD. Supervised topic models. Adv Neural Inf Process Syst. 2010;3:327–32. https://doi.org/10.1109/ICPR.2014.65.
22. Wang C, Blei DM, Li FF. Simultaneous image classification and annotation. In: IEEE conference on computer vision and pattern recognition—CVPR'09. 2009. p. 1903–1910. https://doi.org/10.1109/CVPR.2009.5206800.
23. Li FF, Perona P. A bayesian hierarchical model for learning natural scene categories. In: IEEE computer society conference on computer vision and pattern recognition—CVPR'05, vol 2. 2005. p. 524–531. https://doi.org/10.1109/CVPR.2005.16.
24. Bosch A, Zisserman A, Munoz X. Scene classification via pLSA. Eur Conf Comput Vis. 2006;3954:517–30. https://doi.org/10.1007/11744085_40.
25. Zeng J, Cheung WK, Liu J. Learning topic models by belief propagation. IEEE Trans Pattern Anal Mach Intell. 2013;35(5):1121–34. https://doi.org/10.1109/TPAMI.2012.185.
26. Yuan L, Bin J, Wei Y, Huang F, Hu X, Tan M. Big data aspect-based opinion mining using the slda and hme-lda models. Wirel Commun Mob Comput. 2020;2020:1–19. https://doi.org/10.1155/2020/8869385.
27. Glasserman P, Krstovski K, Laliberte P, Mamaysky H. Choosing news topics to explain stock market returns. In: ACM international conference on AI in finance. 2020. p. 1–8. https://doi.org/10.1145/3383455.3422557.
28. Lu X, Guo Y, Chen J, Wang F. Topic change point detection using a mixed bayesian model. Data Min Knowl Discov. 2021. https://doi.org/10.1007/s10618-021-00804-1.
29. Baburajan V, de Abreu e Silva J, Pereira FC. Open-ended versus closed-ended responses: a comparison study using topic modeling and factor analysis. IEEE Trans Intell Transp Syst. 2021;22(4):2123–32. https://doi.org/10.1109/TITS.2020.3040904.
30. Ma L, Fan S. CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests. BMC Bioinform. 2017;18:1–18. https://doi.org/10.1186/s12859-017-1578-z.
31. Hassan H, Badr A, Abdelhalim MB. Prediction of O-glycosylation sites using random forest and GA-tuned PSO technique. Bioinform Biol Insights. 2015;9:103–9. https://doi.org/10.4137/BBI.S26864.
32. Wei X, Croft WB. LDA-based document models for ad-hoc retrieval. In: Proceedings of the 29th annual international ACM SIGIR. 2006. p. 178–185. https://doi.org/10.1145/1148170.1148204.
33. Panichella A. A systematic comparison of search-based approaches for LDA hyperparameter tuning. Inf Softw Technol. 2021;130:106411. https://doi.org/10.1016/j.infsof.2020.106411.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.