Mach Learn (2017) 106:1787–1815
DOI 10.1007/s10994-017-5638-4

A Bayesian nonparametric model for multi-label learning

Junyu Xuan[1] · Jie Lu[1] · Guangquan Zhang[1] · Richard Yi Da Xu[1] · Xiangfeng Luo[2]

Received: 5 April 2016 / Accepted: 25 March 2017 / Published online: 25 August 2017

Abstract Multi-label learning has become a significant learning paradigm in the past few years due to its broad application scenarios and the ever-increasing number of techniques developed by researchers in this area. Among existing state-of-the-art works, generative statistical models are characterized by their good generalization ability and robustness on large number of labels through learning a low-dimensional label embedding. However, one issue of this branch of models is that the number of dimensions needs to be fixed in advance, which is difficult and inappropriate in many real-world settings. In this paper, we propose a Bayesian nonparametric model to resolve this issue. More specifically, we extend a Gamma-negative binomial process to three levels in order to capture the label-instance-feature structure. Furthermore, a mixing strategy for Gamma processes is designed to account for the multiple labels of an instance. The mixed process also leads to a difficulty in model inference, so an efficient Gibbs sampling inference algorithm is then developed to resolve this difficulty. Experiments on several real-world datasets show the performance of the proposed model on multi-label learning tasks, comparing with three state-of-the-art models from the literature.

Editor: James Cussens.

✉ Jie Lu
Jie.Lu@uts.edu.au

Junyu Xuan
Junyu.Xuan@uts.edu.au

Guangquan Zhang
Guangquan.Zhang@uts.edu.au

Richard Yi Da Xu
Yida.Xu@uts.edu.au

Xiangfeng Luo
luoxf@shu.edu.cn

[1] Faculty of Engineering and Information Technology, University of Technology Sydney, PO Box 123, Broadway, Sydney, NSW 2007, Australia

[2] School of Computer Engineering and Science, Shanghai University, 99 Shangda Road, Shanghai, China

Keywords Multi-label learning · Topic model · Bayesian nonparametric learning

1 Introduction

Multi-label learning (Gibaja and Ventura 2015; Zhang and Zhou 2014; Gao and Zhou 2013) is a significant learning paradigm in which each instance may be assigned more than one label. It has attracted a lot of attentions of not only scholars from research communities but also practitioners from industries in the past few years due to its broad application scenarios (Madjarov et al. 2012). For instance, each academic paper may have more than one author, and learning from this data could help to identify the academic interests of authors and recommend potential collaborators according to their interests (Rosen-Zvi et al. 2004; Xuan et al. 2015b); a patent may be associated with several categories, and automatically assigning large amount of new patents to correct categories could save the costs in human resources and time (Cong and Tong 2008); each gene may be associated with not one but a set of functional classes, and detecting functional classes of new genes could benefit the medicine design (Elisseeff and Weston 2001).

The existing algorithms and models for multi-label learning could be roughly categorized into two types: discriminative ones and generative ones. The generative models learn a joint distribution of data and the latent variables, while discriminative models only learn a conditional distribution of latent variables given data. Comparing with discriminative models, generative ones are characterized by the capability of handling the following situations: (1) the number of labels is large (Rubin et al. 2012); (2) the number of training data is small. Current generative models for multi-label learning are mainly based on topic models (Rai et al. 2015; Rubin et al. 2012), which learn a low-dimensional label embedding (Rai et al. 2015). It means that the labels and instances could be represented by a relatively low-dimensional vector and each dimension of vectors is seen as a topic.

One problem of existing generative models for multi-label learning is that the hidden topic number needs to be fixed in advance. This number is normally chosen with domain knowledge. After fixing the number of topics, Dirichlet, Multinomial, and other distributions could be adopted as the building blocks for generative models. However, discovering an appropriate number is very difficult and sometimes unrealistic for many real-world applications. This may also lead to overfitting when there are too many topics so that topics are relatively specific and do not generalise well to unseen observations; underfitting is the opposite case when there are too few topics so unrelated observations are assigned together to the same topic (Dai and Storkey 2015). A number of methods can be used to choose the number of topics, such as cross-validation techniques (Griffiths and Steyvers 2004), but it is slow because the algorithm has to be restarted a number of times and then choosing the best one (Griffiths and Steyvers 2004; Dai and Storkey 2015). Bayesian nonparametric learning (Hjort et al. 2010; Gershman and Blei 2012) has emerged as an elegant way to handle this problem.

In this paper, we propose a Bayesian nonparametric model for multi-label learning without the requirement of fixing the topic number in advance. Instead of using fixed-dimensional distributions, stochastic processes are used: to be specific, Gamma-negative binomial process (Zhou and Carin 2015) is extended to three levels for capturing the hierarchical structure: label-instance-feature. In this model, each instance is assigned with a Gamma process (Ferguson 1973) to express the mapping relation between this instance with the hidden topics instead of a vector with a fixed dimension. This Gamma process can be simply consid-

ered as an infinite discrete distribution, and is parameterized by a base measure (another Gamma process) that denotes the mapping relation between labels with hidden topics. However, an instance normally has multiple labels in multi-label learning paradigm, so we assign an instance a mixed Gamma process that is from all the Gamma processes of the labels of this instance. Furthermore, introducing mixed Gamma process will lead to intricacies in terms of model inference. Therefore, an efficient Gibbs sampler with closed-form conditional distributions is developed for the proposed model. Experiments on the three multi-label learning tasks with public datasets show the performance of our model comparing existing comparative algorithms or models from the state-of-the-art research literatures.

The main contributions of this paper are:

- a new Bayesian nonparametric model for multi-label learning without the requirement of fixing topic number in advance that is needed by the traditional generative models for multi-label learning;
- theoretical and empirical expectation analysis of the topic number from the proposed mixed Gamma-negative binomial process for understanding the behavior and sensitivity of the process under different parameters;
- an efficient Gibbs sampling inference algorithm for getting the solution of the proposed model which overcomes the inference difficulty brought by the mixing operation in the proposed model.

The rest of this paper is organized as follows. Section 2 briefly reviews related work. Section 3 describes some preliminary knowledge. The mixed Gamma-negative binomial process model is proposed in Sect. 4 with its Gibbs sampling inference algorithm and expectation analysis. Section 5 presents experimental results on three multi-label learning tasks using real-world datasets. Finally, Sect. 6 concludes this study with a discussion on future work.

2 Related work

This section reviews the related work of this study, which is composed of two parts: The first part is about the multi-label learning based on the generative models; and the second part is about Bayesian nonparametric learning.

2.1 Generative models for multi-label learning

It is commonly believed that the mixture model proposed in Mccallum (1999) is the first generative model for multi-label learning, which assigns each label a word distribution and a multi-label document is assumed to be generated according to the word distributions of its labels. This idea is similar with the subsequent topic models (Blei et al. 2003; Xuan et al. 2015a) that are Bayesian models with fixed-dimensional probability distributions. They are originally designed for unsupervised text mining task which aims to discover hidden topics (i.e., word distributions) in the text corpus. Due to their powerful representation and good extendibility, they have been successfully applied to many research areas, including multi-label learning.

One category of using topic model idea for multi-label learning is to directly replace topics in Latent Dirichlet Allocation (LDA) (Blei et al. 2003) by labels, such as Labeled LDA (Ramage et al. 2009) and Flat-LDA (Rubin et al. 2012). Prior-LDA (Rubin et al. 2012) is further proposed to account for the label frequency differences within a corpus through

introducing a label sampling step by multinomial distribution. However, the dependency between the labels is not considered, which is resolved by the Dependency-LDA (Rubin et al. 2012) later. Parametric Mixture Models (Ueda and Saito 2002) are also proposed to capture the pairwise label correlation. More intrinsic correlations among multiple labels are exploited by a model: Labelled Four-level Pachinko Allocation Model (Ma et al. 2012), which is verified with better performance than Labeled LDA (Ramage et al. 2009).

Another category is to assign each label a topic distribution instead of a word distribution, such as Author topic Model (Steyvers et al. 2004; Rosen-Zvi et al. 2010) and Emotion Topic Model (Bao et al. 2012). Each label is first associated with a topic distribution, and each topic is further associated with a word distribution. The generation of a document is split into two stages: (1) generating a topic according to its labels; (2) generating a word according to the drawn topic. CoL model (Wang et al. 2008) also extends this idea with additional label and word correlation learning ability.

To summarize, in spite of the verified success in the multi-label learning of the above models, they all have an issue that the number of topics needs to be fixed in advance. In this paper, we propose a Bayesian nonparametric model to address this issue.

2.2 Bayesian nonparametric learning

Bayesian nonparametric learning (Nguyen and Wu 2015; Nguyen et al. 2013) is a key approach for learning the number of mixtures in a mixture model (also known as model selection problem). Without predefining the number of mixtures, this number is supposed to be inferred from the data, i.e., let the data speak. The idea of Bayesian nonparametric learning is to use stochastic processes to replace traditional fixed-dimensional probability distributions, such as Multinomial, Poisson, and Dirichlet distributions. In order to avoid the limitation associated with fixed dimensions, Multinomial Process (MP), Poisson Process (PP) (Iwata et al. 2013) and Dirichlet Process (DP) (Ferguson 1973) are used to replace former distributions because of their infinite property. The merit of these stochastic processes is that they let the data determine the number of factors (topics, in text mining). DP is a good alternative for the models with Dirichlet distribution as the prior. Many probabilistic models with fixed dimensions have been extended to the infinite ones by the help of stochastic processes: Gaussian Mixture Model (GMM) is extended to Infinite Gaussian Mixture Model (IGMM) (Rasmussen 1999; Ma et al. 2014) using DP; Hidden Markov Model is extended with infinite number of hidden states using Hierarchial Dirichlet Process (HDP) (Teh et al. 2006; Wulsin et al. 2014). Through the posterior inference (i.e., Markov chain Monte Carlo (MCMC) (Neal 2000)), the number of the mixtures can be inferred. Although HDP can model the data with three or more levels, it cannot be directly adopted for the multi-label learning task. The reason is that there is a mixing relationship between authors and documents which cannot be modeled by HDP. Similarly, Partially Labeled Topic Models (PLTM) (Ramage et al. 2011) also cannot be adopted for our problem. Other popular processes including beta process (Hjort 1990), Gamma process, Poisson process, multinomial process, negative binomial process (NBP) (Zhou and Carin 2015; Broderick et al. 2015) have also been successfully used in the machine learning communities recently.

To summarize, Bayesian nonparametric learning (Buntine and Mishra 2014) has been successfully used to extend many finite models and applied to many real-world applications. However, to the best of our knowledge, existing state-of-the-art works cannot be used for multi-label learning. This paper addresses this shortcoming by proposing a mixed Gamma negative binomial process.

3 Preliminary knowledge

This section briefly introduces related concepts which will be used as the building blocks for our proposed model in the following section. To help understanding these concepts, we take the *author-document-word* as an example of multi-label learning throughout this paper where *authors* are seen as *labels*; *documents* are seen as *instances*; *words* are seen as *features*. Several important notations used throughout this paper are summarized in Table 1.

3.1 Gamma process

A Gamma process $GaP(c, H)$ (Ferguson 1973; Roychowdhury and Kulis 2014) is a stochastic process parameterized by a base (shape) measure H and concentration (scale) parameter

Table 1 Notations used in this paper

Notation	Description
Θ	A measurable space
R^+	The set of positive real number
Z^+	The set of positive integer number
D	Number of documents
A	Number of authors
V	Number of different words
K	Number of topics
AD	Author-document mapping matrix
DN	Document-word mapping matrix
A_d	Number of authors of document d
N_d	Number of words of document d
θ_k	Topic k
Γ_0	A global random measure from a Gamma process
$r_{0,k}$	The global weight of topic k
Γ_d	A random measure from a Gamma process for document d
$r_{d,k}$	The weight of topic k in document d (the interest of d on k)
Γ_a	A random measure from a Gamma process for author a
$r_{a,k}$	The weight of topic k in author a (the interest of a on k)
Γ_a^d	The mixed measure of measures of all authors who write d
$r_{a,k}^d$	The average weight of topic k in all author a who write document d
X	A random measure from a Negative binomial process
n_k	Number of words assigned to topic k
X_d	A random measure for document d from a negative binomial process
$n_{d,k}$	Number of words assigned to topic k in document d
$n_{a,k}$	Number of words assigned to topic k and author a
$n_{d,k}^a$	Number of words assigned to topic k and author a in document d
$z_{d,n}$	The topic index assigned to word n in document d
$i_{d,n}$	The author index assigned to word n in document d
ϖ_a^d	The weight of author a in document d

c. Let $\Gamma = \{(r_k, \theta_k)\}_{k=1}^{\infty}$ be a random realization of a Gamma process in the product space $R^+ \times \Theta$. Then, we have

$$\begin{aligned} \Gamma &\sim GaP(c, H) \\ &= \sum_{k=1}^{\infty} r_k \delta_{\theta_k} \end{aligned} \tag{1}$$

where δ_{θ_k} is a Dirac measure parameterized by θ_k (i.e., $\delta_{\theta_k}(\hat{\theta}) = 1$ if $\hat{\theta} = \theta_k$; 0, otherwise); r_k satisfies an improper Gamma distribution $Gamma(0, c)$; and $\theta_k \sim H$. Γ also corresponds to a complete random measure (Kingman 1992; Zhou and Carin 2015). When Γ is assigned to a document, we can understand θ_k as a topic (i.e., V-dimensional normalized vector) and r_k is the (unnormalized) weight of this topic in this document although the summation of $\{r_k\}_k^{\infty}$ may not be equal to one.

3.2 Negative binomial process

A negative binomial process $NBP(p, \Gamma_0)$ (Zhou and Carin 2015) is also a stochastic process parameterized by a base measure Γ_0 and p. Similar with the Gamma process, a realization of negative binomial process $X = \{(n_k, \theta_k)\}_{k=1}^{\infty}$ is also a set of points in product space $Z^+ \times \Theta$. Then, we have

$$\begin{aligned} X &\sim NBP(p, \Gamma_0) \\ &= \sum_{k=1}^{\infty} n_k \delta_{\theta_k} \end{aligned} \tag{2}$$

where $\{n_k\}$ are integers so negative binomial process is normally used as the likelihood of counting models (Broderick et al. 2015); and $\theta_k \sim \Gamma_0$. Note that if Γ_0 is a continuous measure, the probability that two θ_k are equal is zero; if Γ_0 is a discrete measure, say $\Gamma_0 = \sum_{k=1}^{\infty} \delta_{\widetilde{\theta}_k}$, θ_k can only take the value from $\{\widetilde{\theta}_k\}_{k=1}^{\infty}$. Compared with Poisson process which is another alternative for the counting model, negative binomial process has a better variance-to-mean ratio (VMR) and the overdispersion level (Simon 1960; Zhou and Carin 2015). When X is assigned to a document, θ_k can be understood as a topic and n_k can be understand as the number of words in this document assigned to topic θ_k.

3.3 Gamma-negative binomial process

Normally, negative binomial process is used as the likelihood part of a Bayesian nonparametric model. Analogous to a negative binomial distribution $x \sim NB(r, p)$ which has two parameters: $r > 0$ and $p \in [0, 1]$, there are two kinds of priors for the parameters of a negative binomial process: one is Gamma process for Γ_0 as shown in Eq. (1) (Zhou and Carin 2015) ; the other is the Beta process for p (Broderick et al. 2015). In this paper, we use the Gamma process prior. A Gamma-negative binomial process model is proposed in (Zhou and Carin 2015) as shown in Fig. 1 and it can be represented as,

$$\begin{aligned} \Gamma_0 &\sim GaP(c_0, H) \\ X_d &\sim NBP(p_d, \Gamma_0) \end{aligned} \tag{3}$$

where p_d is a real-valued parameter within [0, 1] and the base measure of the negative binomial process Γ_0 is a random measure from a Gamma process. X_d is for a document, and this hierarchial form makes the documents share a same base measure Γ_0. This Gamma-negative

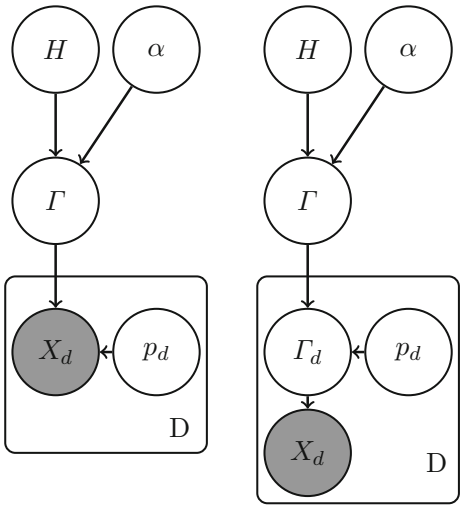

Fig. 1 Gamma-negative binomial process model. The *left subfigure* is related to Eq. (3) and the *right hand* part is related to Eq. (4)

binomial process can be (in distribution) equivalently augmented as Gamma-Gamma-Poisson process,

$$\begin{aligned} \Gamma_0 &\sim GaP(c_0, H) \\ \Gamma_d &\sim GaP\left(\frac{1-p_d}{p_d}, \Gamma_0\right) \\ X_d &\sim PP(\Gamma_d) \end{aligned} \tag{4}$$

where $PP(\Gamma_d)$ is a Poisson process with parameter Γ_d. This augmentation is useful for the closed-form model inference algorithm design.

4 Mixed gamma-negative binomial processes

In this section, we first propose a mixed Gamma-negative binomial processes model (MGNBP) while *author-document-word* is still taken as an example to explain why this model could be used for multi-label learning in Sect. 4.1; We then introduce a Gibbs sampler to inference the proposed model in Sect. 4.2; A significant property, i.e., expectation of topic number, is theoretically and empirically analyzed in Sect. 4.3.

4.1 Model description

Consider the Gamma-negative binomial process model in Eqs. (3) and (4) again: despite its success, this model however is fundamentally the same as the basic topic models, which are used for modeling the data of two level hierarchy: instance-feature (i.e., document-word). Multi-label learning requires to model the data with a three-level hierarchy: label-instance-feature (i.e., author-document-word). So an intuitive idea is to add another Gamma process level to capture the additional label (i.e, author) level based on the Gamma-negative binomial process model in Eq. (4) analogues to the hierarchical mechanism of Hierarchical Dirichlet Process (Teh et al. 2006),

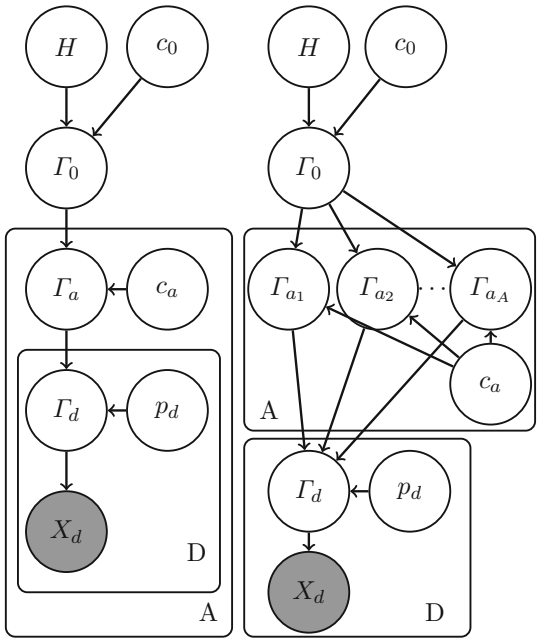

Fig. 2 Gamma-Gamma-Negative Binomial Process Model (*left one*) and Mixed Gamma-Negative Binomial Process Model (*right one*)

$$\begin{aligned} \Gamma_0 &\sim GaP(c_0, H) \\ \Gamma_a &\sim GaP(c_a, \Gamma_0) \\ \Gamma_d &\sim GaP((1-p_d)/p_d, \Gamma_a^d) \\ X_d &\sim PP(\Gamma_d) \end{aligned} \tag{5}$$

where Γ_a is the new added level for the label (i.e, author). We call this model three-level Gamma-negative binomial process model (3GNBP), which is graphically shown in the left subfigure of Fig. 2.

More specifically, the global measure in the 3GNBP model is

$$\Gamma_0 = \sum_{k=1}^{\infty} r_{0,k} \delta_{\theta_k} \tag{6}$$

where $r_{0,k}$ is the global weight of topic θ_k. This global measure defines a set of global topics $\{\theta_k\}_{k=1}^{\infty}$ shared by all documents, and $\{r_{0,k}\}_{k=1}^{\infty}$ indicates the overall "interests" of documents on topics. The number of topics can be potentially infinite and therefore justifies the infinity in the summation. However, since the data is limited, the learned topics will be also limited. Each author a is then assigned a realization of Gamma process parameterized by Γ_0,

$$\Gamma_a = \sum_{k=1}^{\infty} r_{a,k} \delta_{\theta_k} \tag{7}$$

where $r_{a,k}$ is the weight of k-th topic θ_k which is inherited from the global measure Γ_0. $\{r_{a,k}\}_{k=1}^{\infty}$ can be viewed as the "interest" of author a on the topics $\{\theta_k\}_{k=1}^{\infty}$. Similarly to the author, each document is also assigned a realization of Gamma process parameterized by Γ_a,

$$\Gamma_d = \sum_{k=1}^{\infty} r_{d,k}\delta_{\theta_k} \tag{8}$$

where $\{r_{d,k}\}_{k=1}^{\infty}$ is the weight of "interests" of document d on the topics inherited from the global measure Γ_0 again. In the 3GNBP model, the base measure Γ_a for Γ_d is from its author. It can be seen as the 'interest inheritance'. Finally, the likelihood is a realization of Poisson process,

$$X_d = \sum_{k=1}^{\infty} n_{d,k}\delta_{\theta_k} \tag{9}$$

where $n_{d,k}$ is the number of words in document d assigned to topic k.

When applying 3GNBP to multi-label learning, there is a significant issue that each Γ_d could only have one parent Γ_a as its base measure which means that each instance is with one and only one label (i.e., a document could only have one author). Therefore, the intuitive idea of 3GNBP cannot be used for the multi-label learning. In order to resolve this issue, our innovative idea is to combine all the Gamma processes of all authors of a document together by

$$\Gamma_a^d = \varpi_{a_1}^d \Gamma_{a_1} + \varpi_{a_2}^d \Gamma_{a_2} + \cdots + \varpi_{a_{A_d}}^d \Gamma_{a_{A_d}} \tag{10}$$

where A_d is the number of labels of an instance (i.e., authors of document d); $\varpi_{a_1}^d$ is the weight of label a_1 on instance d and $\sum_a \varpi_a^d = 1$ (i.e., the contribution of author a_1 to document d) which is given a Dirichlet prior $Dir(\eta)$; and Γ_a^d is the mixed prior for Γ_d. Note that the plus here is element-wise because each Γ_a is with a countably infinite number of components. This element-wise plus action is reasonable because the components of each Γ_a are countable and they are all with same discrete base measure Γ_0. We can see the mixed Gamma process Γ_a^d as the "mixed interest" of all the authors of a document. This document has "inherited" the interests on the topics from the "mixed interest" not from the interest of an author. Through this way, the multiple labels of an instance could be modeled. To summarize, our proposed Mixed Gamma-Negative Binomial Processes Model (MGNBP) is as follows

$$\begin{aligned}
&\Gamma_0 \sim GaP(c_0, H)\\
&\Gamma_a \sim GaP(c_a, \Gamma_0)\\
&\Gamma_a^d = \varpi_{a_1}^d \Gamma_{a_1} + \varpi_{a_2}^d \Gamma_{a_2} + \cdots + \varpi_{a_{A_d}}^d \Gamma_{a_{A_d}}\\
&\Gamma_d \sim GaP((1-p_d)/p_d, \Gamma_a^d)\\
&X_d \sim PP(\Gamma_d)
\end{aligned}$$

and its graphical representation is shown in the right subfigure of Fig. 2.

4.2 Model inference

It is difficult to perform posterior inference under infinite mixtures, and a commonly work-around solution in Bayesian nonparametric learning is to use a truncation method (Fox et al. 2011; Blei et al. 2010). Truncation method is widely accepted, which uses a relatively big $K^{\dagger}$ as the (potential) maximum number of topics. Under the truncation, the model can be expressed below as a good approximation to the infinite model,

$$
\begin{aligned}
\theta_{1:K^\dagger} &\sim \frac{1}{\gamma_0} H \\
\gamma_0 &\sim Gamma(e_0, 1/f_0) \\
r_{0,k}|\gamma_0, c_0 &\sim Gamma(\gamma_0/K^\dagger, 1/c_0) \\
r_{a,k}|r_0, c_a &\sim Gamma(r_{0,k}, 1/c_a) \\
p_d &\sim Beta(a_0, b_0) \\
r^d_{a,k} &= \varpi^d_{a_1} r_{a_1,k} + \varpi^d_{a_2} r_{a_2,k} + \cdots \varpi^d_{a_{A_d}} r_{a_{A_d},K^\dagger} \\
r_{d,k}|r_a, p_d &\sim Gamma(r^d_{a,k}, p_d/(1-p_d)) \\
n_{d,k} &\sim Pois(r_{d,k}) \\
N_d &= \sum_{k=1}^{K^\dagger} n_{d,k}
\end{aligned}
$$

and $n_{d,k}$ could also be equivalently (in distribution) generated as follow

$$
\begin{aligned}
z_{d,n} &\sim Multi(r_{d,1}/\sum r_d, \ldots, r_{d,K^\dagger}/\sum r_d) \\
w_{d,n} &\sim \theta_{z_{d,n}} \\
n_{d,k} &= \sum_n \delta_{(z_{d,n}==k)}
\end{aligned}
$$

where $Pois()$ denotes a Poisson distribution; $Multi()$ denotes a multinomial distribution; $\gamma_0 = \int dH$ is the total mass of measure H; and the parameters are given the appropriate priors. Here, H is a V-dimensional Dirichlet distribution, and each θ is a topic that is a V-dimensional vector.

The difficult part of the inference for this model is the mixed part Γ^d_a or r^d_a. Since r^d_a is the mixed value, it is hard to infer the posterior of r_a through its likelihood. In order to resolve this issue, we firstly introduce the Additive Property of the negative binomial distribution: *If X_i follows a negative binomial distribution with parameters r_i and p and if the various X_i are independent, then $\sum X_i$ follows a negative binomial distribution with parameters $\sum r_i$ and p.*

In MGNBP model, we have

$$
\begin{aligned}
r_{d,k}|\{r_a\}, p_d &\sim Gamma(r^d_{a,k}, p_d/(1-p_d)) \\
n_{d,k} &\sim Pois(r_{d,k})
\end{aligned}
\tag{11}
$$

which are (in distribution) equal to

$$
n_{d,k} \sim NB(r^d_{a,k}, p_d) \tag{12}
$$

and according to Additive Property of negative binomial distribution, it is further (in distribution) equal to

$$
\begin{aligned}
n^a_{d,k} &\sim NB\left(\varpi^d_a \cdot r_{a,k}, p_d\right) \\
n_{d,k} &= \sum_a n^a_{d,k}
\end{aligned}
\tag{13}
$$

where $NB()$ denotes a negative binomial distribution and $\{n^a_{d,k}\}$ are independent with each others.

We have split $n_{d,k}$ the number of words assigned to topic k in document d into a number A_d of independent variables $\{n^a_{d,k}\}$. Here, $n^a_{d,k}$ denotes the number of words assigned to topic k from author a in document d. From Eq. (13), we can see that we have obtained the likelihood part of the r_a, so we can update/inference the r_a using n^a_d. Introducing the auxiliary variables $\{n^a_{d,k}\}$ helps us resolve the difficult inference problem brought by the mixed Gamma process. Note that the independence between the elements of $\{n^a_{d,k}\}$ is very important, which facilitates us to update each $n^a_{d,k}$ independently.

According to the relationship between the negative binomial distribution and the Gamma-Poisson distribution, for each $n^a_{d,k}$, we have:

$$\begin{aligned} & n^a_{d,k} \sim NB(\varpi^d_a \cdot r_{a,k}, p_d) \\ \Longrightarrow & r^a_{d,k} \sim Gamma(\varpi^d_a \cdot r_{a,k}, p_d/(1-p_d)),\ n^a_{d,k} \sim Pois(r^a_{d,k}) \end{aligned} \tag{14}$$

We want to highlight that $r^a_{d,k}$ is different from $r^d_{a,k}$: $r^d_{a,k}$ is the mixed Gamma process of multiple author Gamma processes Γ_a of Gamma process Γ_d of document d and $r^a_{d,k}$ is the interest of document d on topic k inherited from author a.

Due to the non-conjugacy of Gamma distribution and negative binomial distribution, it is difficult to update r_a with a Gamma prior. In order to make the inference with only close-formed conditional distributions, we use the following result on the negative binomial process,

Theorem 1 (Zhou and Carin 2015) *If X follows a negative binomial distribution $X \sim NB(r, p)$ with parameters r and p, then X can also be generated from a compound Poisson distribution as*

$$X = \sum_{t=1}^{l} u_t, u_t \overset{i.i.d}{\sim} Log(p), l \sim poiss\,(-rln(1-p)) \tag{15}$$

where $Log()$ is a Logarithmic distribution. Furthermore, this Poisson-logarithmic bivariate count distribution, $p(X, l)$, can be expressed as

$$X \sim NB(r, p),\ l \sim CRT(X, r) \tag{16}$$

where $CRT()$ denotes a Chinese Restaurant Table distribution, and its definition and sampling can be found in (Zhou and Carin 2015).

With Theorem 1, the Eq. (14) is also equal to

$$\begin{aligned} n^a_{d,k} \sim NB(\varpi^d_a \cdot r_{a,k}, p_d) \Longrightarrow & n^a_{d,k} \sim \sum_{1}^{l^a_{d,k}} log(p_d), l^a_{d,k} \sim Pois(-\varpi^d_a r_{a,k} ln(1-p_d)) \\ \Longrightarrow & l^a_{d,k} \sim CRT(n^a_{d,k}, \varpi^d_a r_{a,k}), n^a_{d,k} \sim NB(\varpi^d_a r_{a,k}, p_d) \end{aligned} \tag{17}$$

Finally, we can update all $n^a_{d,k}$ by,

$$\begin{aligned} (n^{a_1}_{d,k_1}, \ldots, n^{a_A}_{d,K} | N_d) \sim & Mult\left(N_d, \frac{\varpi^d_{a_1} r^{a_1}_{d,k_1}}{r_d}, \ldots, \frac{\varpi^d_{a_{A_d}} r^{a_A}_{d,K}}{r_d}\right) \\ r_d = & \sum_a \sum_k \varpi^d_a \cdot r^a_{d,k} \end{aligned} \tag{18}$$

and for each word n in a document d, we can assign it to a topic k and author a by

$$\begin{aligned} p(z_{d,n} = k, i_{d,n} = a) &\propto \frac{\varpi_a^d r_{d,k}^a}{r_d} \\ n_{d,k} &= \sum_n \delta(z_{d,n} = k) \\ n_{a,k} &= \sum_d \sum_n \delta(z_{d,n} = k \ \& \ i_{d,n} = a) \end{aligned} \tag{19}$$

where $z_{d,n}$ is the topic index assigned to word n in document d.

With these changes of variables, the original model is re-formulated as,

$$\begin{aligned} \gamma_0 &\sim Gamma(e_0, 1/f_0) \\ r_{0,k}|\gamma_0, c_0 &\sim Gamma(\gamma_0/K^\dagger, 1/c_0) \\ p_d &\sim beta(a_{d,0}, b_{d,0}) \\ r_{a,k}|r_0, c_a &\sim Gamma(r_{0,k}, 1/c_a) \\ r_{a,k}^d &= \varpi_{a_1}^d r_{a_1,k} + \varpi_{a_2}^d r_{a_2,k} + \cdots \varpi_{a_{A_d}}^d r_{a_{A_d},K^\dagger} \\ r_{d,k}|r_a, p_d &\sim Gamma(r_{a,k}^d, p_d/(1-p_d)) \\ r_{d,k}^a &\sim Gamma(\varpi_a^d r_{a,k}, p_d/(1-p_d)), \quad a \in A^d \\ z_{d,n}^a &\sim Category\left(\frac{\varpi_a^d r_{d,k}^a}{r_d}, \cdots\right) \\ n_{d,k} &= \sum_n \delta(z_{d,n} = k) \\ n_{a,k} &= \sum_d \sum_n \delta(z_{d,n} = k \ \& \ i_{d,n} = a) \\ n_{d,k}^a &= \sum_n \delta(z_{d,n} = k \ \& \ i_{d,n} = a) \\ N_d &= \sum_n \sum_a z_{d,n}^a \end{aligned} \tag{20}$$

where $Category()$ denotes a Category distribution; A^d is the set of associated authors of document d; and $|A^d| = A_d$ is the cardinality of A^d.

In the following, a Gibbs sampling algorithm (Andrieu et al. 2003) is designed for the posterior inference and all the conditional distributions are listed in the Appendix. We can see from these conditional distributions that all of them are closed-form which is very easy to update and implement. The whole procedure is summarized in Algorithm 1. Note that after we obtain all the samples of the posterior $p(\theta, r_a, r_d, r_0, z_{d,n}^a, p_d, \gamma_0, n_{d,k}^a|\cdots)$ of latent variables and remove the burn-in stage, we firstly identify the topic number with largest frequency as the K_{real}, and then find the sample with largest likelihood and $K = K_{real}$ from these samples. The output of Gibbs sampler are the latent variables θ, r_a and r_d in this sample.

4.3 Model analysis

A distinguishing characteristic of Bayesian nonparametric model is that the number of the factors/topics to be learned is not specified in advance. Roughly speaking, Bayesian non-

Algorithm 1: Gibbs Sampler for MGNBP

Input: D, A, N, AD, DN
Output: K_{real}, $\{\theta\}$, $\{r_a\}$, $\{r_d\}$
initialization;
while $iter \leq max_{iter}$ **do**
  for $d = 1; d \leq D$ **do**
    for $n = 1; n \leq N_d$ **do**
      Update $z_{d,n}$ and $i_{d,n}$ by Eq. (32);
    for $a = 1; a \leq A_d$ **do**
      Update $r^a_{d,k}$ by Eq. (33);
      Update $l^a_{d,k}$ by Eq. (34);
    Update $r_{d,k}$ and p_d by Eq. (35);
  for $a = 1; a \leq A$ **do**
    Update $r_{a,k}$ by Eq. (36);
    Update $l_{a,k}$ by Eq. (37);
  Update $r_{0,k}$ by Eq. (38);
  Update l'_k by Eq. (40);
  Update γ_0 by Eq. (41);
  Update θ by Eq. (43);
  $iter++$;
Identify K_{real};
Select the sample with largest likelihood and $K = K_{real}$;
return $\{\theta\}$, $\{r_a\}$, $\{r_d\}$;

parametric model could be simply seen as a prior for the this number. Conditioned on the observed data, we could determine how many factors/topics are needed. It would be interesting to investigate the prior expectation of the factors/topics number under our defined model. We give the following result,

Theorem 2 *Given D instances, A labels, and their mapping AD, the expected factor number from the MGNBP is*

$$\int_{r_0} \left(1 - \prod_a \left[\frac{c_a}{c_a - \sum_{d:AD[a,d]>0} \varpi_a^d \ln(1-p_d)}\right]^{r_0}\right) \cdot \frac{\gamma_0 \cdot \exp(-c_0 \cdot r_0)}{r_0} \cdot dr_0 \tag{21}$$

and when a truncation level $K^\dagger$ is applied, the expected factor number is

$$K^\dagger \left(1 - \left[\frac{c_0}{c_0 - \sum_a \log \frac{c_a}{c_a - \sum_{d:AD[a,d]>0} \varpi_a^d \ln(1-p_d)}}\right]^{\frac{\gamma_0}{K^\dagger}}\right) \tag{22}$$

where γ_0, c_0, c_a and p_d are four parameters of the MGNBP.

Proof We first introduce the following theorem of a completely random measure,

Theorem 3 (Kingman 1992) ***Campbell's Theorem*** *Let Π be a Poisson process on Θ with mean measure μ, and let $f : \Theta \to \mathbb{R}$ be measurable. Then the sum*

$$\sum = \sum_{Y \in \Pi} f(Y) \tag{23}$$

is absolutely convergent with probability if and only if

$$\int_\Theta \min(|f(y)|, 1)\mu(dy) < \infty \tag{24}$$

If this condition holds, the expectation

$$\mathbb{E}\left[\sum\right] = \int_{\Theta} f(y)\mu(dy) \tag{25}$$

exists if and only if the integral converges.

Since the proposed MGNBP is a completely random measure, we can utilize the above theorem to compute the expectation of sum of its variables. We define a random variable,

$$\mathbb{X}_k = \mathbb{1}\left(\sum_{d=1}^{D} \sum_{AD[a,d]=1} C_{d,k}^{a} > 0\right) \tag{26}$$

which equals to 1 if the factor k is used; 0, otherwise. So the expected factor number is $\mathbb{E}[\sum_k \mathbb{X}_k]$. Then, according to the Theorem 3,

$$\begin{aligned}
\mathbb{E}\left[\sum_k \mathbb{X}_k\right] &= \mathbb{E}\left[\mathbb{E}\left[\sum_k \mathbb{X}_k | r_{0,k}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[\sum_k \mathbb{X}_k | \{r_{a,k}\}\right] | r_{0,k}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\sum_k \mathbb{E}\left[\mathbb{X}_k | \{r_{a,k}\}\right] | r_{0,k}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\sum_k \left(1 - \prod_{d=1}^{D} \prod_{a:AD[a,d]>0} (1-p_b)^{\varpi_a^d r_{a,k}}\right) | r_{0,k}\right]\right] \\
&= \mathbb{E}\left[\sum_k \mathbb{E}\left[\left(1 - \prod_{d=1}^{D} \prod_{a:AD[a,d]>0} (1-p_d)^{\varpi_a^d r_{a,k}}\right) | r_{0,k}\right]\right] \\
&= \mathbb{E}\left[\sum_k \int_{r_{1,k}} \cdots \int_{r_{A,k}} \left(1 - \prod_{d=1}^{D} \prod_{a:AD[a,d]>0} (1-p_d)^{\varpi_a^d r_{a,k}}\right)\right. \\
&\quad \left. \cdot \left\{\prod_{a=1}^{A} \frac{{c_a}^{r_{0,k}}}{\Gamma(r_{0,k})} {r_{a,k}}^{r_{0,k}-1} \exp(-c_a \cdot r_{a,k})\right\} dr_{1,k} \cdots dr_{A,k}\right] \\
&= \mathbb{E}\left[\sum_k \left(1 - \prod_a \left[\frac{c_a}{c_a - \sum_{d:AD[a,d]>0} \varpi_a^d \ln(1-p_d)}\right]^{r_{0,k}}\right)\right] \\
&= \int_{r_0} \left(1 - \prod_a \left[\frac{c_a}{c_a - \sum_{d:AD[a,d]>0} \varpi_a^d \ln(1-p_d)}\right]^{r_0}\right) \cdot \nu_{GaP}(r_0) \cdot dr_0 \\
&= \int_{r_0} \left(1 - \prod_a \left[\frac{c_a}{c_a - \sum_{d:AD[a,d]>0} \varpi_a^d \ln(1-p_d)}\right]^{r_0}\right) \\
&\quad \cdot \frac{\gamma_0 \cdot \exp(-c_0 \cdot r_0)}{r_0} \cdot dr_0
\end{aligned}$$

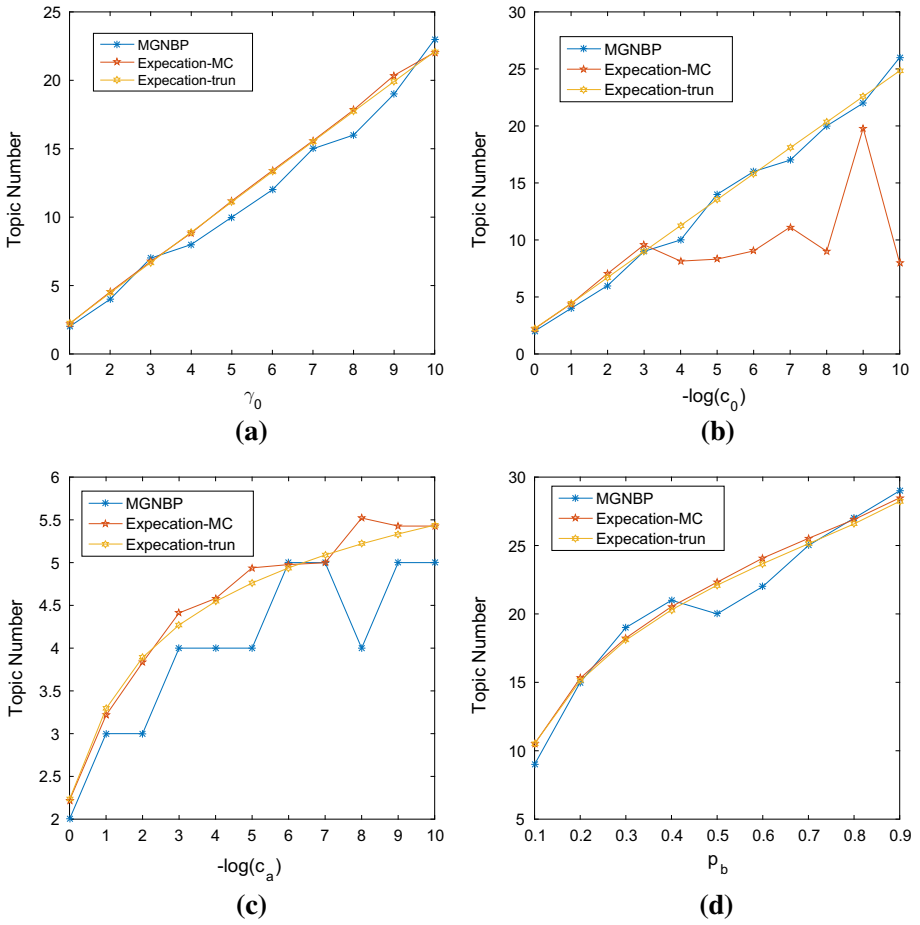

Fig. 3 The comparisons between expected and empirical factor numbers of MGNBP under different parameters: γ_0, c_0, c_a and p_d. Note that the x-axes of c_0 and c_a are in negative (base-10) log space, **a** parameter γ_0, **b** parameter c_0, **c** parameter c_a, **d** parameter p_d

This integral cannot be easily solved. An approximate method for this integral computation is Monte Carlo. If we apply a truncation level $K^\dagger$, the expectation is

$$\mathbb{E}\left[\sum_k^{K^\dagger}\mathbb{X}_k\right] = \int_{r_{0,1}\cdots r_{0,K^\dagger}} \sum_k^{K^\dagger}\left(1-\prod_a\left[\frac{c_a}{c_a-\sum_{d:AD[a,d]>0}\varpi_a^d\ln(1-p_d)}\right]^{r_{0,k}}\right) \cdot \prod_{k=1}^{K^\dagger}\frac{{c_0}^{\gamma_0/K^\dagger}}{\Gamma(\gamma_0/K^\dagger)}{r_{0,k}}^{\gamma_0/K^\dagger-1}\exp(-c_0 r_{0,k})dr_{0,1}\cdots dr_{0,K^\dagger}$$

$$= \sum_k \left\{ 1 - \left[\frac{c_0}{c_0 - \sum_a \log \frac{c_a}{c_a - \sum_{d:AD[a,d]>0} \varpi_a^d \ln(1-p_d)}} \right]^{\frac{\gamma_0}{K^\dagger}} \right\}$$

$$= K^\dagger \left(1 - \left[\frac{c_0}{c_0 - \sum_a \log \frac{c_a}{c_a - \sum_{d:AD[a,d]>0} \varpi_a^d \ln(1-p_d)}} \right]^{\frac{\gamma_0}{K^\dagger}} \right)$$

The theorem is proved. □

Our above theoretical result is also supported by simulation results, summarized in Fig. 3. At first, we set $A = 10$, $D = 20$, and the mapping relations between labels and instances are randomly generated. We simulate the model with the above setting and different values of parameters, and then compare the empirical factor number and the theoretical factor number from Theorem 2. The default values of them are: $\gamma_0 = 1$, $c_0 = 1$, $c_a = 1$, $p_d = 0.5$, and ϖ_a^d of instances are equal for all labels. When investigating one parameter, the other three will be fixed as the default values. Four subfigures in Fig. 3 denote the changing of factor number as the changing of four parameters of the model, respectively. In each subfigure, *Expectation-MC* denotes the Monte Carlo approximation of the expectation of the factor number from Eq. (21); *Expectation-trun* denotes the truncation-based approximation of the expectation of the factor number from Eq. (22) (Note that the implementation of the MGNBP is based on the truncation $K^\dagger = 1000$); the x-axes of c_0 and c_a are in negative (base-10) log space. From these results, we can see that the theoretical factor number is very close to the empirical factor number, so this verified our results on the expected factor number of the MGNBP. The trends of the empirical and expected factor number with parameters γ_0 and p_d are very close with each others. For the parameter c_a, the trends are also close; for the parameter c_0, *Expectation-MC* is a little away from the others as the increasing of the value of c_0. These subfigures do not only verify the above theoretical result, they also show the sensitivity of the model to the parameters.

5 Experiments

In this section, we evaluate the performance of the proposed Mixed Gamma-Negative Binomial Processes Model (MGNBP) on three multi-label learning tasks: *author topic modeling*, and *clinical free text labeling*, and *protein classification*, and the proposed model is also compared with six state-of-the-art models or algorithms using the public datasets of these tasks.

5.1 Datasets

The datasets used in the experiments are:

– **NIPS papers**[1] This dataset contains papers from the NIPS conferences between 1987 and 1999. This dataset is a structure: *author-paper-word*. It contains 1740 papers with 2037 authors, a total of 2,301,375 word tokens and a vocabulary size of 13,649 unique words. More descriptions can be found in Steyvers et al. (2004);

[1] http://www.datalab.uci.edu/author-topic/NIPs.htm.

Table 2 Statistics of datasets

Datasets	Label number	Instance number	Feature number
NIPS	2037	1740	13,649
DBLP	28,702	28,569	11,771
Clinical	45	978	1449
Protein	27	662	1185

- **DBLP papers**[2] The abstracts and authors of papers are extracted through DBLP interface from four areas: database, data mining, information retrieval and artificial intelligence. More descriptions can be found in Deng et al. (2011);
- **Clinical free texts**[3] This dataset is a structure: *label-text-feature*. There are 45 labels (like ICD-9-CM codes) and 645 (training) / 333 (testing) data with 1,449 features. More descriptions can be found in Pestian et al. (2007).
- **Proteins** (see footnote 3) This dataset is a structure: *class-protein-feature*. There are 27 categories for these protein sequences, e.g., PDOC50007 (a class of hydrolases), and 463 (training) / 199 (testing) data with 1,185 features, i.e., Prosite access numbers. More descriptions can be found in Diplaris et al. (2005).

The statistics of datasets are shown in Table 2.

5.2 Author-topic modeling task

Since the proposed model is motivated to resolve the multi-label learning problem in Introduction, the author-paper data as a kind of multi-label data is appropriate to evaluation of the efficiency of the proposed model on multi-label learning. We use an author's distribution over topics to characterize this author (i.e., the author research interest), and the dimension of this distribution is not fixed in advance but learned from the data owing to the Bayesian nonparametric learning technique. Based on the distribution over topics of each author, there are a number of practical applications, for example, 1) Collaborator Recommendation. People with similar research interest may have the potential to be collaborators. We can recommend researcher A to researcher B by simply evaluating the similarity of their interest vectors. 2) Author Disambiguation. Some researchers may have exactly same name on their scientific papers, so it is hard to distinguish them through name. Based on the learned interests of researchers with the same name, we can identify the real author of a paper through comparing/differencing the content of this paper with the two authors' research interests. Since these practical applications are both based on the output of the proposed model: author interests, we only evaluate the proposed model on the author interest learning in the manuscript. If the author interests are more accurately learned, the performance of the model on the above practical applications will apparently be better as well.

5.2.1 Experiment setting

For the first two datasets, we randomly select some documents as training data and test data. The number of selected training documents is around 1000, and the number of test documents

[2] http://www.cs.uiuc.edu/~hbdeng/data/kdd2011.htm.

[3] http://mulan.sourceforge.net/datasets-mlc.html.

is about 30 percent of the number of training documents. The requirement of selections is: the training and test documents must share some authors and words. This requirement makes sure the learned topics and authors' interests can be used to predict the test documents. We compare the proposed model with **Author-Topic Model (ATM)**(Steyvers et al. 2004)[4] which could be seen as a generative model for multi-label learning using fixed dimensional distributions. Another comparative model is **Disjoint Author-Document Topic model (DADT)**(Seroussi et al. 2014) which models documents and authors using two separate sets of topics.

The first evaluation metric is *Perplexity* which is widely used in language modeling to assess the predictive power of a model (Steyvers et al. 2004; Blei et al. 2003). The perplexity is a measure of how surprising the words in the test documents are from the model's perspective and can be calculated by

$$Perplexity = \exp\left(-\sum_{d}\sum_{k} p(w_d|\theta_k)p(\theta_k|a_d)\right) \tag{27}$$

where a_d is the authors of test document d. The smaller the value of perplexity is, the better the predictive ability of a model has. Since we use the same test documents for different models, the normalization is not considered because it does not influence the model comparisons.

The second evaluation metric is *logLikelihood* of training data,

$$logLikelihood = \sum_{d} \log p(w_d|\theta, r_a, r_d) \tag{28}$$

This is a measure of the probability of the training documents under the learned latent variables θ, r_a and r_d. It can be understood as 'how the model fits the training data'. The larger the value of likelihood is, the better a model fits the training data. Likelihood in Eq. (28) is to show the ability to model the training data and Perplexity in Eq. (27) is to show the ability to predict the test data. We think these two commonly-adopted and complementary metrics are sufficient for the model comparison.

Another evaluation metric *AuthorP* is designed for evaluating the author prediction based on learned topics.

$$AuthorP = \frac{1}{N_d^t}\sum_{d}^{N_d^t} < w_{a,d}, w_d >, \quad w_{a,d} = \sum_{a}^{A_d} \varpi_d^a r_a\theta \tag{29}$$

where N_d^t is the number of test documents, w_d is the word-vector of document d, $w_{a,d}$ is the average word-vector of all authors of document d, and $r_a\theta$ is the word-vector of author a. The word-vector of an author, in fact, indicates the probability of this author writing different words, so we use the similarity between average word-vector of all authors of a document and the word-vector of this document to evaluate the possibility of these authors writing this document. It appears that the larger *AuthorP* is, the better the model is. Note that *Perplexity* is designed for evaluating the document prediction based on learned topics, so they are different.

5.2.2 Result analysis

For the *DBLP* dataset, the comparative results between MGNBP and ATM are shown in Fig. 4. Each row of the Fig. 4 denotes a group of DBLP dataset. The left subfigures show the comparison on the data log-likelihood. Here, we adjust different active topic numbers

[4] http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm.

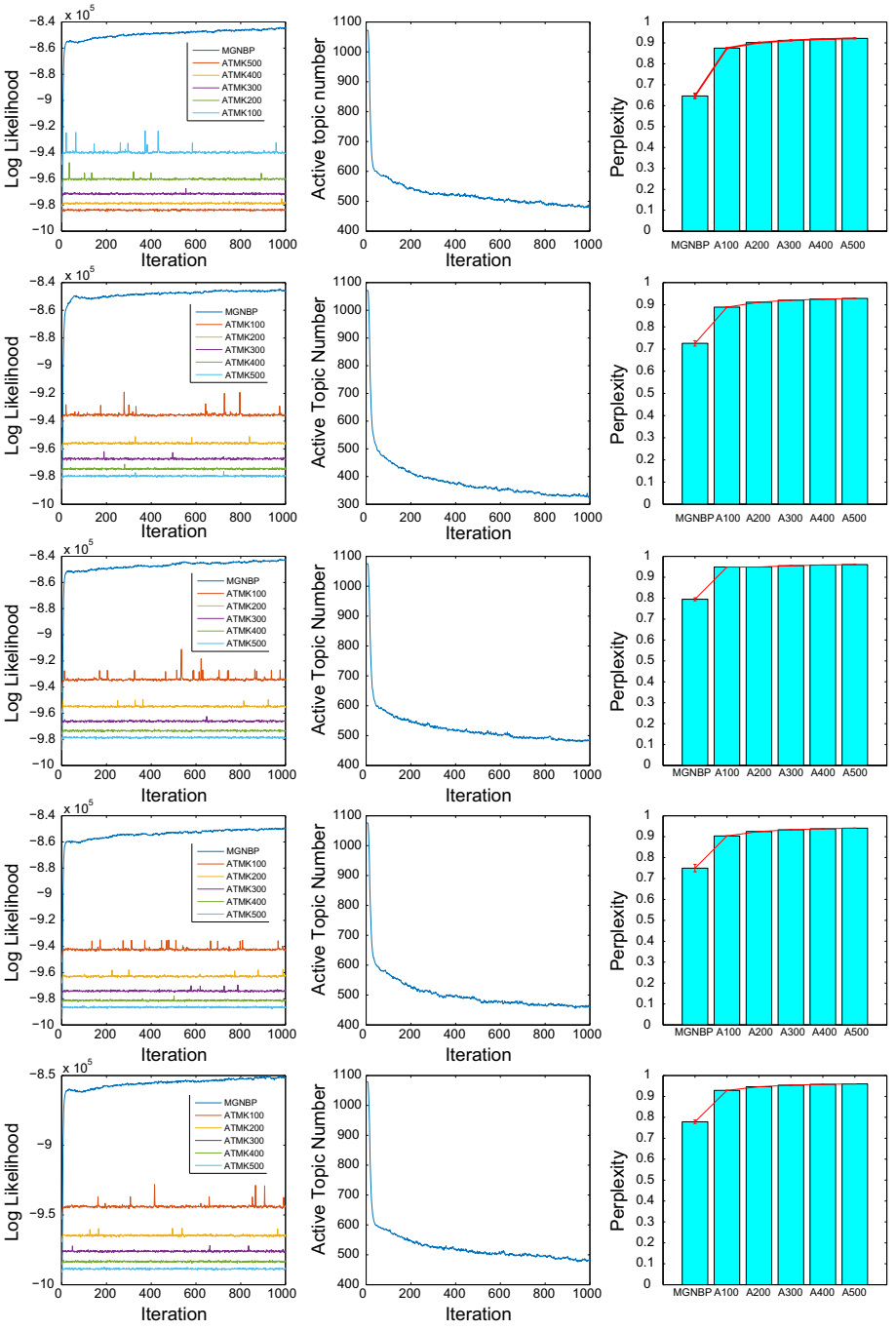

Fig. 4 Results from MGNBP and ATM with different (predefined) topic numbers on five groups of DBLP dataset. *Each row* denotes a group. In each row, the *left subfigure* shows the Log-likelihoods comparison; The *middle subfigure* shows the change of active topic number of MGNBP during the iteration; the *right subfigure* shows the perplexity comparison

for the ATM, including $K = 100$, $K = 200$, $K = 300$, $K = 400$ and $K = 500$. From these subfigures, the proposed MGNBP model (The hyper-parameters are set as following by experiences for the rest of this section: $a_0 = 1$, $b_0 = 1$, $e_0 = 1$, $f_0 = 1$, $c_0 = 1$ and $c_a = 1$) outperforms the ATM on different preset topic numbers. It means that MGNBP fits the training documents better than the ATM, and, more importantly, MGNBP does not depend on the domain knowledge to predefine the active topic number, making the method widely applicable. The middle subfigures in Fig. 4 indicate the changing of active topics during the iteration of the MGNBP (The number of active topics is set as the number of training documents at the initialization step of the model). These curves show that the number of active topics dramatically drops down at the burn-in stage of the sampling, and begins to stabilize after about 200 iterations. Since the documents are different in content but similar in numbers amongst the groups, the learned topic number differs slightly amongst each others. These numbers are: group 1: $K = 519$; group 2: $K = 332$; group 3: $K = 493$; group 4: $K = 465$; group 5: $K = 504$. We also compare the performances of two models (MGNBP and ATM) on the test documents prediction using perplexity in Eq. (27). Since the training and test documents share some authors, we can compute the perplexity of the test documents according to the learned topics and authors' interests on them. At each step of iterations, the perplexity of test documents is computed using the latent variables, i.e., $\{\theta\}$, $\{r_a\}$ and $\{r_d\}$, at this iteration. The results are shown in right subfigures of Fig. 4. In each subfigure, the first bar denotes the mean of perplexities of all iterations except the burn-in stage ($1 \sim 200$ iterations) of the proposed model MGNBP and the others denote ATM with different (predefined) topic numbers. The standard deviations are also shown in the subfigures. The proposed model gets the best performance (smallest perplexity). The standard deviation of MGNBP is relatively bigger than ATM. The reason is because the number of active topics will change during the iteration but it will not change in ATM, so in theory, the random-walk space of Gibbs sampler of MGNBP should be larger than that of ATM. Even with this relatively larger standard deviation, the mean of perplexity of MGNBP is smaller than ATM.

For the *NIPS* dataset, the comparative results between MGNBP and ATM are shown in Fig. 5. Same with the *DBLP* dataset, the log likelihoods of MGNBP and ATM with different predefined active topic numbers are shown in the left side of the Fig. 5. Unsurprisingly, the subfigures in the middle column show the convergence of MGNBP (group 1: 367; group 2: 529; group 3: 354). Specially, we found that the log-likelihoods of ATM increases when topic number decreases. Therefore, we have compared with ATM with only two (the minimum number) topics as shown in the left subfigures in Fig. 5. It can be seen that the proposed MGNBP model also gets larger log likelihood and smaller perplexity when compared with ATM except the case where ATM is set to have 10 topics in group 2. Even so, the ATM in group 2 with 10 topics has almost same performance with MGNBP on the Log-likelihood of training documents. Moreover, we can see that it takes 800 iterations to reach this stability for the ATM with 10 topics, but MGNBP only takes fewer than 50 iterations to reach the same stability. It is worth mentioning that ATM achieves its best Perplexity when only two topics are involved. The reason is that the Perplexity in Eq. (27) inherently prefers smaller K due to its definition/equation in this paper. This is not only unique to our work which uses Gamma-Nonnegative Binomial Processes to obtain an optimal K. The comparisons made in the previous topic model which uses fixed K also have this phenomenon.

We also compare the proposed model with DADT and ATM on the author prediction using *NIPS* dataset. Since DADT and ATM are fixed-dimensional probabilistic models, we feed them the following dimensionality candidates: $< 10, 20, 30, \ldots, 100 >$. The results are listed in Table 3. It can be seen from the table that the author prediction results from DADT and ATM will fluctuate with the change of dimensionality but the result from MGNBP does

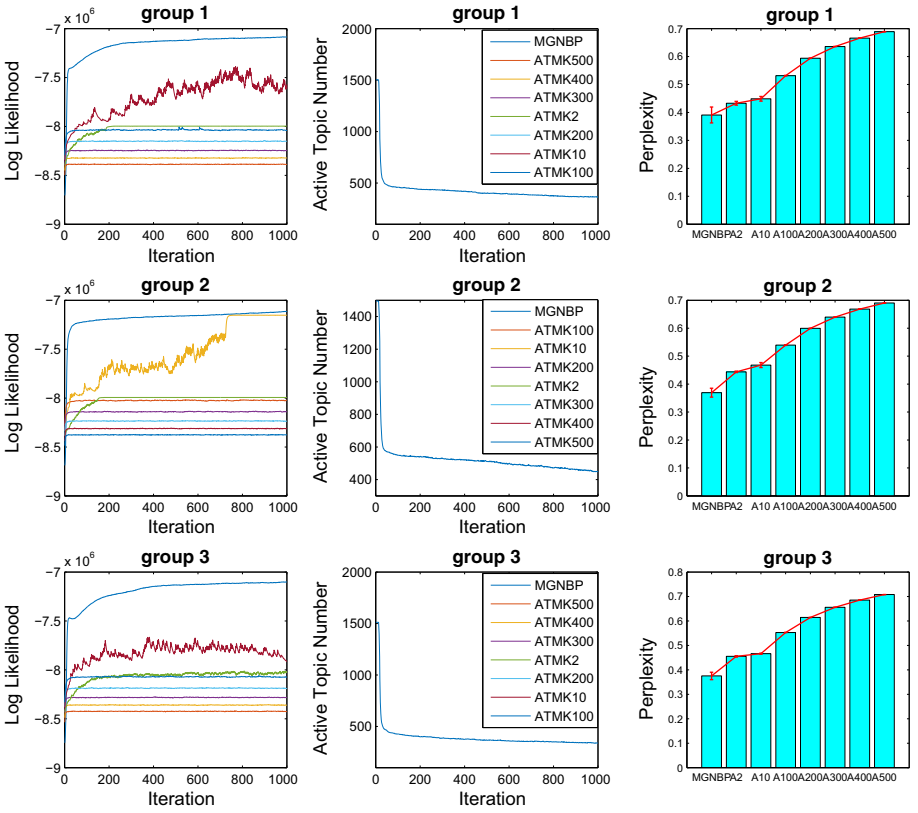

Fig. 5 Results from MGNBP and ATM with different (predefined) topic numbers on three groups of NIPS dataset. *Each row* denotes a group. In each row, the *left subfigure* shows the Log-likelihoods comparison; The *middle subfigure* shows the change of active topic number of MGNBP during the iteration of Gibbs sampler; the *right subfigure* shows the perplexity comparison

Table 3 Comparisons on Author prediction with *NIPS* dataset

Models		DADT	ATM	MGNBP
Dimensionality	*K=10*	0.8436	0.5749	0.8431
	K=20	0.8434	0.5704	
	K=30	0.8430	0.5733	
	K=40	0.8430	0.5694	
	K=50	0.8430	0.5688	
	K=60	0.8430	0.5713	
	K=70	0.8413	0.5709	
	K=80	0.8413	0.5685	
	K=90	0.8359	0.5741	
	K=100	0.8352	0.5685	

Terms are highlighted in italics to link the dataset

not. We can draw the conclusion that MGNBP could achieve better performance than ATM and comparative performance with DADT but MGNBP is not with additional prerequisite.

5.3 Clinical free text labeling task

Clinical free texts are primary data about patients. Manually labeling these clinical free texts is a challenge due to the expensive cost of labor. For example, the cost of adding labels like ICD-9-CM to clinical free texts and repairing associated errors is approximately 25 billion per year in the US (Pestian et al. 2007). Since each text may be associated with more than one code, multi-label learning could be adopted to accomplish this task, i.e, automatically label clinical free texts at an very low cost.

5.3.1 Experiment setting

The comparative models for this task are **LEAD** (Zhang and Zhang 2010) and **LIFT** (Zhang and Wu 2015),[5] which are both deterministic models based on Support Vector Machine. Comparing with LEAD and LIFT, the proposed model is a generative model, class models which normally have better generalizing ability on the unseen data compared with deterministic models, especially with small datasets.

In multi-label learning area, it is commonly accepted that ranking labels for the test data is as valuable as predicting labels, so many multi-label classification models or algorithms return a probability vector for a test datapoint (with each dimension representing a label) rather than predicting the labels for a test datapoint (Zhang and Zhou 2014). In order to evaluate the returned label probability vector, ranking-based evaluation metrics have been proposed in the literature, including *Oneerror*, *Coverage*, *Rankingloss*, *Avgprecision* (Gibaja and Ventura 2015). For *Avgprecision*, the larger the value, the better the performance; For *Oneerror*, *Coverage* and *Rankingloss*, the smaller the value, the better the performance. The core of these metrics is to compute the probability of a test datapoint x_i with a specific label l, i.e., $R(l, x_i)$. Next, we will introduce how to compute this probability using the trained proposed model.

From the proposed model, we could obtain new representations for all labels and (training) instances, i.e., r_a and r_d, which are both K-dimensional vectors. Given a test document, we can obtain its interest r_i as the expectation of its posterior distribution,

$$p(r_i|\cdots) \propto \int_{p_d}\int_{\varpi^d}\left(\prod_n \sum_{z_{i,n}} \sigma_{i,v}\, p(w_{i,n}|\{\theta\}, z_{i,n})\, p(z_{i,n}|r_i)\right) p(r_i|\{r_a\}, p_d, \varpi^d, \cdots)\, p(p_d)\, p(\varpi^d|\eta)\, d(p_d)\, d(\varpi^d) \tag{30}$$

where $\sigma_{i,v} \in [0, 1]$ is the weight of a test datapoint x_i on feature v and V is the total number of features of test datapoint x_i. With the interest of labels (i.e., $\{r_a\}$) and a test datapoint (i.e., r_i), their similarity is computed as the probability of x_i with l by

$$R(l, x_i) = < \overrightarrow{r_l}, \overrightarrow{r_i} > \tag{31}$$

where $<, >$ denotes the cosine similarity function. This metric is reasonable because the datapoint is with a label when they have similar interest on the hidden topics, which is consistent with the assumption of the proposed model.

[5] Implementations are both from: http://cse.seu.edu.cn/people/zhangml/Resources.htm.

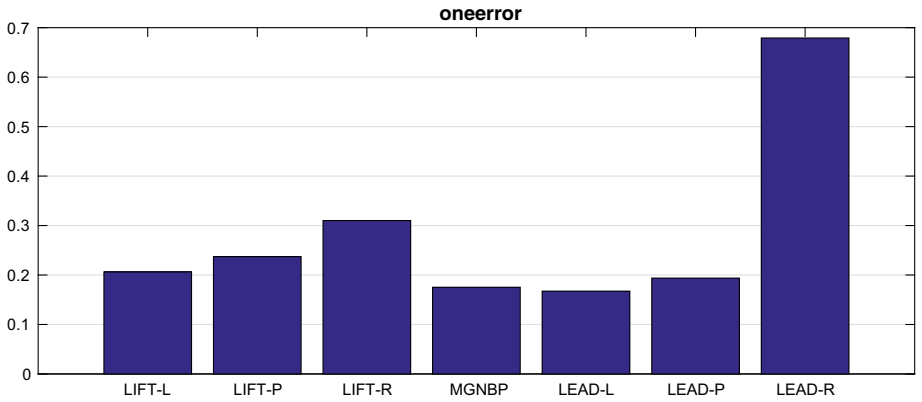

Fig. 6 The comparisons between LIFT, LEAD, and MGNBP on Clinical free text labeling task on *Oneerror* (The smaller the value, the better the performance). 'XX-L', 'XX-P' and 'XX-R' denotes XX model with Linear kernel function, Polynomial kernel function and Radial basis function (RBF) kernel function

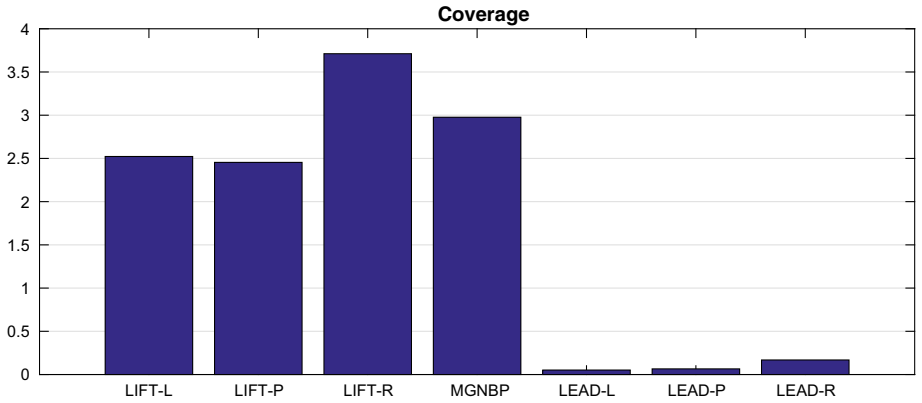

Fig. 7 The comparisons between LIFT, LEAD, and MGNBP on Clinical free text labeling task on *Coverage* (The smaller the value, the better the performance). 'XX-L', 'XX-P' and 'XX-R' denotes XX model with Linear kernel function, Polynomial kernel function and Radial basis function (RBF) kernel function

5.3.2 Results analysis

Since the LEAD and LIFT are SVM-based models, we have compared their different implementations using different kernel functions. 'LIFT-L' denotes LIFT with Linear kernel function; 'LIFT-P' denotes LIFT with Polynomial kernel function (the degree is set as 3); 'LIFT-R' denotes LIFT with radial basis function (RBF) kernel function. The results have been shown in Figs. 6, 7, 8, and 9, which show the results on four evaluation metrics respectively. From these Figures, we can see that MGNBP achieves good performances on *Oneerror* and *Rankingloss*, and it also obtains the comparative performance on *Avgprecision*. For the *Coverage*, LEAD achieves the best performances, and MGNBP is only better than the worst LIFT-R. Four metrics have their own preferences on the classification evaluation. Among the four metrics, *Oneerror* and *Coverage* are like 'variance', and *Rankingloss* and *Avgprecision* are like 'mean'. So the proposed model has better performance on the 'mean' (on average), and at the same time the 'variance' is not very larger than LIFT and LEAD. The reason

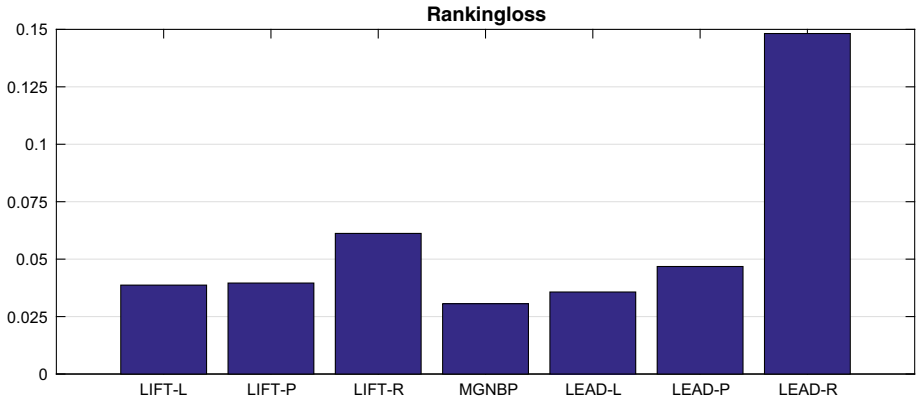

Fig. 8 The comparisons between LIFT, LEAD, and MGNBP on Clinical free text labeling task on *Rankingloss* (The smaller the value, the better the performance). 'XX-L', 'XX-P' and 'XX-R' denotes XX model with Linear kernel function, Polynomial kernel function and Radial basis function (RBF) kernel function

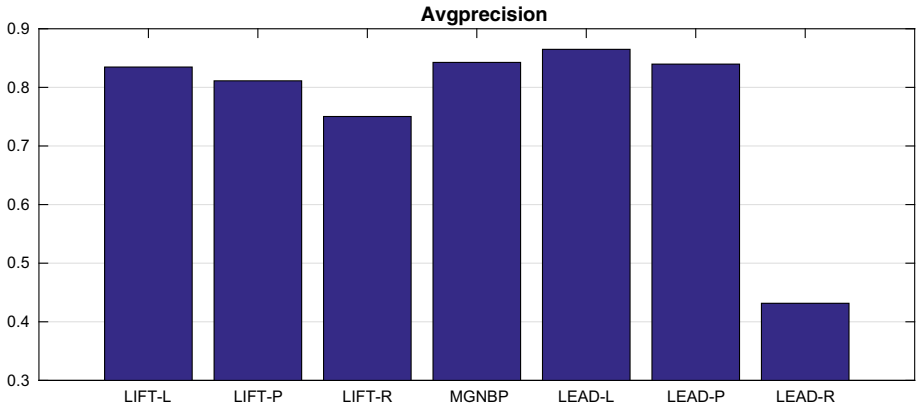

Fig. 9 The comparisons between LIFT, LEAD, and MGNBP on Clinical free text labeling task on *Avgprecision* (The larger the value, the better the performance). 'XX-L', 'XX-P' and 'XX-R' denotes XX model with Linear kernel function, Polynomial kernel function and Radial basis function (RBF) kernel function

may be that the proposed model is based on MCMC, so each run is a sample from the real model distribution and there will be variance during the sampling although the variance has already been decreased by the incorporating of the data. To summarize, MGNBP has better performance than LIFT and LEAD on the labeling task, especially considering the generative model nature of MGNBP.

5.4 Protein classification task

The number of proteins stored in protein databases keeps growing all over the world. These Proteins could be grouped into several families according to their functions from structures, which is valuable for a number of biology applications, e.g., new medicine design. However, not all the proteins have been correctly classified by the researchers because the laboratory

Table 4 Protein classification results by three models

Models	Evaluation metrics			
	Avgprecision	*Coverage*	*Oneerror*	*Rankingloss*
BCS	0.5255 ± 0.4941	10.7774 ± 10.9242	0.0015 ± 0.0024	0.5015 ± 0.5254
BMLPL	0.4668 ± 0.3693	9.2055 ± 10.2758	0.6920 ± 0.3437	0.0188 ± 0.0191
MGNBP	0.9045	1.3719	0.1206	0.0225

experiments are often expensive. Fortunately, the multi-label learning model could be trained to predict the categories of the unlabeled proteins at a low cost.

5.4.1 Experiment setting

The comparative state-of-the-art models for this task are **Bayesian Compressed Sensing (BCS)** (Kapoor et al. 2012) and **Bayesian Multi-label Learning via Positive Labels (BMLPL)** (Rai et al. 2015). Different from LEAD and LIFT in Sect. 5.3, BCS and BMLPL are both based on Bayesian framework so they belong to generative models same with the proposed MGNBP. Unfortunately, two models are both with a low-dimensional embedding, so it needs to predefine the dimensionality for them. In contrast, the proposed model, i.e., MGNBP, does not have this prerequisite. The evaluation metrics, i.e., *Oneerror*, *Coverage*, *Rankingloss*, and *Avgprecision*, used in Sect. 5.3.1 are still adopted in this experiment.

5.4.2 Results analysis

The classification results on *Protein* dataset of three models, i.e., BCS, BMLPL, and MGNBP, are listed in Table 4. This table records the predictions results from three models on four evaluation metrics. As stated before, BCS and BMLPL are two fixed-dimensional Bayesian models, so there will be fluctuations in their results according to different prefixed dimensionality. In this experiment, we run the two models with dimensionality: $\{10, 20, 30, \ldots, 90, 100\}$. Each cell in Table 4 from BCS and BMLPL is composed of two numbers: mean and standard deviation of the results on 10 different dimensionality. In contrast, MGNBP does not need the dimensionality as input, so the cell in Table 4 from MGNBP only contains one number. It appears that avoiding the result fluctuation is one advantage of MGNBP compared to BCS and BMLPL. We can see from the numbers in the table that the fluctuations of BCS and BMLPL on four metrics are all significant. It means that the dimensionality setting can significantly affect the classification results of BCS and BMLPL. When facing new data without any prior knowledge, selecting an appropriate dimensionality would be very difficult. MGNBP achieves the best performance on *Avgprecision* and *Coverage*. On *Oneerror*, BCS is the best one, and MGNBP is much better than BMLPL. On *Rankingloss*, MGNBP achieves a little worse but comparable performance with BMLPL, and it is much better than BCS. To sum up, we conclude that without the prerequisite of setting dimensionality, MGNBP can still achieve good performance on this task.

6 Conclusions and further study

We have developed a Bayesian nonparametric model for multi-label learning that can automatically learn a latent factor/topic embedding for both labels and instances without the need

of fixing factor/topic number that is a common issue for most existing generative models for multi-label learning. In the proposed model, we have extended Gamma-negative binomial process into three layers with additional Gamma process layer to capture the three-layer hierarchy: label-instance-feature. Furthermore, a mixing strategy has been designed to combine the information of different labels for an instance which accounts for the multi-label setting. The expected topic number has been theoretically and empirically analyzed. The comparative experiments with three state-of-the-art algorithms and models in literature on two real-world multi-label learning tasks have demonstrated the effectiveness of the proposed model.

Another further study is to design a variational inference algorithm for the proposed model because current Gibbs sampling-based inference cannot scale well to the big data.

Acknowledgements Research work reported in this paper was partly supported by the Australian Research Council (ARC) under discovery Grants DP140101366 and DP150101645.

Appendix: Conditional distributions for MCMC

Sampling z

$$p(z_{d,n} = k, i_{d,n} = a|\cdots) \propto \theta_{k,n} \cdot \varpi_a^d r_{d,k}^a \tag{32}$$

Sampling r_d^a

$$p(r_{d,k}^a|\cdots) \propto Gamma(\varpi_a^d r_{a,k} + n_{d,k}^a, p_d) \tag{33}$$

Sampling l_d^a

$$p(l_{d,k}^a|\cdots) \propto CRT\left(n_{d,k}^a, \varpi_a^d r_{a,k}\right) \tag{34}$$

Sampling p_d

$$\begin{aligned} r_{a,k}^d &= \varpi_{a_1}^d r_{a_1,k} + \varpi_{a_2}^d r_{a_2,k} + \cdots \\ p(p_d|\cdots) &\propto Beta\left(a_0 + \sum_k n_{d,k}, b_0 + \sum_k r_{a,k}^d\right) \\ p(r_{d,k}|\cdots) &\propto Gamma(r_{a,k}^d + n_{d,k}, p_d) \end{aligned} \tag{35}$$

Sampling r_a

$$p(r_{a,k}|\cdots) \propto Gamma\left(r_{0,k} + \sum_{d \text{ with } a} l_{d,k}^a, \frac{1}{c_a - \sum_{d \text{ with } a} \varpi_a^d \cdot \ln(1 - p_d)}\right) \tag{36}$$

Sampling l_a

$$p(l_{a,k}|\cdots) \propto CRT\left(\sum_{d \text{ with } a} l_{d,k}^a, r_{0,k}\right) \tag{37}$$

Sampling $r_{0,k}$

$$p(r_{0,k}|\cdots) \propto Gamma\left(\gamma_0/K^\dagger + \sum_a l_{a,k}, \frac{1}{c_0 - \sum_a \ln(1 - p_a)}\right) \tag{38}$$

where

$$p_a = \frac{-\sum_{d \text{ with } a} \varpi_a^d ln(1 - p_d)}{c_a - \sum_{d \text{ with } a} \varpi_a^d ln(1 - p_d)} \tag{39}$$

Sampling l'_k

$$p(l'_k|\cdots) \propto CRT\left(\sum_a l_{a,k}, \gamma_0/K^{\dagger}\right) \tag{40}$$

Sampling γ_0

$$p(\gamma_0|\cdots) \propto Gamma\left(e_0 + \sum_k l'_k, \frac{1}{f_0 - \ln(1-p')}\right) \tag{41}$$

where

$$p' = \frac{-\sum_a ln(1-p_a)}{c_0 - \sum_a ln(1-p_a)} \tag{42}$$

Sampling θ_k

$$p(\theta_k|\cdots) \propto H(\theta_k) \prod_d \theta_{z_{d,n}=k,n} \tag{43}$$

Sampling ϖ^d

$$p(\varpi^d|\cdots) \propto Dir(\varpi^d; \eta) \prod_{k=1}^{K} Gamma(r_{d,k}|r_a, p_d, \{r_a\}, \varpi^d) \tag{44}$$

The weights ϖ^d in each document, which could be seen as additional output of the model, could be learned from data. Note that the parameter η will be also updated during the inference, which represents the overall weight of each label in documents. For a test document, we set it as the expectation of its conditional posterior distribution:

$$p(\varpi^d|\cdots) \propto Dir(\varpi^d; \eta) \int_{r_d} \prod_n \sum_{z_{d,n}} Category(w_{d,n}|\{\theta\}, z_{d,n}) Multi(z_{d,n}|r_d)$$
$$\int_{p_d} \prod_k Gamma(r_{d,k}|\{r_a\}, p_d, \varpi^d) Beta(p_d) d(p_d) d(r_d)$$

where $w_{d,n}$ is the n^{th} word of a test document d, $\{\theta\}$ are the learned topics, and $\{r_a\}$ are learned authors' interests on topics.

References

Andrieu, C., de Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning*, *50*(1), 5–43.

Bao, S., Xu, S., Zhang, L., Yan, R., Su, Z., Han, D., et al. (2012). Mining social emotions from affective text. *IEEE Transactions on Knowledge and Data Engineering*, *24*(9), 1658–1670.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, *57*(2), 7.

Broderick, T., Mackey, L., Paisley, J., & Jordan, M. (2015). Combinatorial clustering and the Beta negative binomial process. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(2), 290–306.

Buntine, W,L., & Mishra, S. (2014). Experiments with non-parametric topic models. In *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '14* (pp. 881–890). New York, NY: ACM.

Cong, H., & Tong, L. H. (2008). Grouping of TRIZ inventive principles to facilitate automatic patent classification. *Expert Systems with Applications*, *34*(1), 788–795.

Dai, A. M., & Storkey, A. J. (2015). The supervised hierarchical Dirichlet process. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(2), 243–255.

Deng, H., Han, J., Zhao, B., Yu, Y., & Lin, C.X. (2011). Probabilistic topic models with biased propagation on heterogeneous information networks. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '11* (pp. 1271–1279). New York, NY: ACM.

Diplaris, S., Tsoumakas, G., Mitkas, P,A., & Vlahavas, I. (2005). Protein classification with multiple algorithms. In *Panhellenic conference on informatics, PCI '05* (pp. 448–456). Volos, Greece: Springer.

Elisseeff, A., & Weston, J. (2001). A kernel method for multi-labelled classification. In *Advances in neural information processing systems, NIPS '14* (pp. 681–687). Vancouver, British Columbia, Canada: MIT Press.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, *1*(2), 209–230.

Fox, E., Sudderth, E., Jordan, M., & Willsky, A. (2011). Bayesian nonparametric inference of wwitching dynamic linear models. *IEEE Transactions on Signal Processing*, *59*(4), 1569–1585.

Gao, W., & Zhou, Z. H. (2013). On the consistency of multi-label learning. *Artificial Intelligence*, *199–200*, 22–44.

Gershman, S. J., & Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*(1), 1–12.

Gibaja, E., & Ventura, S. (2015). A tutorial on multilabel learning. *ACM Computing Surveys*, *47*(3), 52:1–52:38.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5228–5235.

Hjort, N. L. (1990). Nonparametric Bayes estimators based on Beta processes in models for life history data. *Annals of Statistics*, *18*(3), 1259–1294.

Hjort, N. L., Holmes, C., Müller, P., & Walker, S. G. (2010). *Bayesian nonparametrics* (Vol. 28). Cambridge: Cambridge University Press.

Iwata, T., Shah, A., & Ghahramani, Z. (2013). Discovering latent influence in online social activities via shared cascade poisson processes. In *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '13* (pp 266–274). New York, NY: ACM.

Kapoor, A., Viswanathan, R., & Jain, P. (2012). Multilabel classification using Bayesian compressed sensing. In *Advances in neural information processing systems, NIPS '25* (pp. 2654–2662). Lake Tahoe, Nevada: ACM.

Kingman, J. F. C. (1992). *Poisson processes* (Vol. 3). Oxford: Oxford university press.

Ma, H., Chen, E., Xu, L., & Xiong, H. (2012). Capturing correlations of multiple labels: A generative probabilistic model for multi-label learning. *Neurocomputing*, *92*, 116–123. (data Mining Applications and Case Study).

Ma, Z., Rana, P. K., Taghia, J., Flierl, M., & Leijon, A. (2014). Bayesian estimation of Dirichlet mixture model with variational inference. *Pattern Recognition*, *47*(9), 3143–3157.

Madjarov, G., Kocev, D., Gjorgjevikj, D., & Deroski, S. (2012). An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, *45*(9), 3084–3104. (best Papers of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'2011)).

Mccallum, A.K. (1999). Multi-label text classification with a mixture model trained by EM. In *Association for the Advancement of Artificial Intelligence Workshop on Text Learning*. Orlando, Florida: AAAI Workshop, AAAI Press.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *9*(2), 249–265.

Nguyen, T. M., & Wu, Q. J. (2015). A non-parametric Bayesian model for bounded data. *Pattern Recognition*, *48*(6), 2084–2095.

Nguyen, V. A., Boyd-Graber, J., Resnik, P., Cai, D. A., Midberry, J. E., & Wang, Y. (2013). Modeling topic control to detect influence in conversations using nonparametric topic models. *Machine Learning*, *95*(3), 381–421.

Pestian, J.P., Brew, C., Matykiewicz, P., Hovermale, D.J., Johnson, N., Cohen, K.B., & Duch, W. (2007). A shared task involving multi-label classification of clinical free text. In *Proceedings of the workshop on BioNLP 2007: Biological, translational, and clinical language processing Bio, NLP '07* (pp. 97–104). Stroudsburg, PA; Association for Computational Linguistics.

Rai, P., Hu, C., Henao, R., & Carin, L. (2015). Large-scale Bayesian multi-label learning via topic-based label embeddings. In Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (Eds.) *Advances in neural information processing systems, NIPS '28* (pp. 3204–3212). Montreal, Quebec: Curran Associates, Inc.

Ramage, D., Hall, D., Nallapati, R., & Manning, C.D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 conference on empirical methods*

in natural language processing: Volume 1 - Volume 1, EMNLP '09 (pp. 248–256). Stroudsburg, PA: Association for Computational Linguistics.

Ramage, D., Manning, C.D., & Dumais. S. (2011). Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, KDD'11* (pp. 457–465). New York, NY: ACM.

Rasmussen, C.E. (1999). The infinite Gaussian mixture model. In *Advances in neural information processing systems, NIPS '12* (pp. 554–560). Denver, CO: ACM.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents UAI '04 (pp. 487–494). Arlington, Virginia: AUAI Press.

Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., & Steyvers, M. (2010). Learning author-topic models from text corpora. *ACM Transactions on Information Systems*, *28*(1), 4:1–4:38.

Roychowdhury, A.,& Kulis, B. (2014). Gamma processes,stick-breaking, and variational inference. arXiv preprint arXiv:1410.1068.

Rubin, T. N., Chambers, A., Smyth, P., & Steyvers, M. (2012). Statistical topic models for multi-label document classification. *Machine Learning*, *88*(1–2), 157–208.

Seroussi, Y., Zukerman, I., & Bohnert, F. (2014). Authorship attribution with topic models. *Computational Linguistics*, *40*(2), 269–310.

Simon, L.J. (1960). The negative binomial and Poisson distributions compared. In *Proceedings of Casualty Actuarial Society, PCAS '60* (vol 47, pp. 20–24).

Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, KDD '04* (pp. 306–315). New York, NY: ACM.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, *101*(476),

Ueda, N., & Saito, K. (2002). Parametric mixture models for multi-labeled text. In *Advances in neural information processing systems, NIPS '15* (pp. 721–728). Vancouver, British Columbia: MIT Press.

Wang, H., Huang, M., & Zhu, X. (2008). A generative probabilistic model for multi-label classification. In *IEEE international conference on data mining, ICDM '08* (pp. 628–637). Pisa: IEEE.

Wulsin, D. F., Fox, E. B., & Litt, B. (2014). Modeling the complex dynamics and changing correlations of epileptic events. *Artificial Intelligence*, *216*, 55–75.

Xuan, J., Lu, J., Zhang, G., & Luo, X. (2015a). Topic model for graph mining. *IEEE Transactions on Cybernetics*, *45*(12), 2792–2803.

Xuan, J., Lu, J., Zhang, G., Xu, R.Y.D., & Luo, X. (2015b). Infinite author topic model based on mixed Gamma-negative binomial process. In *IEEE international conference on data mining, ICDM '15* (pp. 489–498). Atlantic City, NJ: IEEE.

Zhang, M. L., & Wu, L. (2015). LIFT: Multi-label learning with label-specific features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(1), 107–120.

Zhang, M.L., Zhang, K. (2010). Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '10* (pp. 999–1008). New York, NY: ACM.

Zhang, M. L., & Zhou, Z. H. (2014). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *26*(8), 1819–1837.

Zhou, M., & Carin, L. (2015). Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(2), 307–320.