CSI 5155 Assignment3

Yu Sun(8472921)

November 2020

The paper I chose is "A Bayesian nonparametric model for multi-label learning"[1], which mainly focus on the problem of multi-label learning.

1 Research problem

For multi-label problem, especially in text categorization, the generative statistical models are often used because of their good generalization ability. Generally speaking, the models are mixture models, representing the mixture distributions within an overall population. But the number of mixtures is need to be predefined, like the topic number in topic model(Latent Dirichlet allocation). So, the paper proposed a method, which can find the number automatically for the generative statistical models of multi-label problem.

2 Importance(Application and relevance)

Defining an appropriate number is very important but difficult for many real-world applications. If the number is big, it may lead to model over-fitting, while small one may cause under-fitting. One way to solve this problem is cross-validation but the process is time-consuming especially in the large data set. So, learning the distribution number automatically form the data is useful and important.

3 Novelty

The main novelty of this paper is that the authors propose a Bayesian nonparametric model for multi-label learning without the requirement of the distribution number in advance. Besides, the authors also do the theoretical and empirical expectation analysis to prove the model efficiency for the model.

4 Methodology

To summarize, the author used the Bayesian nonparametric learning and stochastic processes to deal with the traditional fixed distribution number. Based on

analysing the limitation of Hierarchial Dirichlet Process for mutil-label process, the author propose a mixed Gamma-negative binomial process, which used multiple Gamma processes not only a hierarchical one to represent different labels. Then regarding to inferring the mixed part of the proposed model, the author introduce the Additive Property of negative binomial distribution and a complete Gibbs sampling strategy is designed.

5 Results

For evaluating the model performance, the author did the analysis on three multi-label learning tasks, including author topic modeling task, clinical free text labeling task and protein classification task. With respect to each task, two state-of-the-art models were used to compared with the proposed model.

Task	Data Sets	Metrics	SOTA models
author topic modeling	NIPS, DBLP	Perplexity, LogLikelihood, AuthorP	DADT, ATM
clinical free text labeling	Clinical free texts	Oneerror, Coverage, Rankingloss, Avgprecision	LEAD, LIFT
protein classification	Proteins	Oneerror, Coverage, Rankingloss, Avgprecision	BCS, BMLPL

Table 1: The Model Evaluation

For task 1, the author concluded that the proposed model could achieve better performance than ATM and comparative performance with DADT but MGNBP is not with additional prerequisite.

For task 2, the the proposed model has better performance on the Rankingloss and Avgprecision, but Oneerror and Coverage are larger than LIFT and LEAD. Among the metrics, Oneerror and Coverage are like 'variance', and Rankingloss and Avgprecision are like 'mean'.

For task 3, the proposed model can achieve the best performance on Avgprecision and Coverage. On Oneerror, BCS is the best one, and the proposed model is much better than BMLPL. On Rankingloss, the proposed model achieves a little worse but comparable performance with BMLPL, and it is much better than BCS.

To sum up, the authors concluded that without the prerequisite of setting distribution number, the proposed model can still achieve comparative performance on three tasks.

6 Limitations

One limitation I think the author did not tell anything about the training time when used the proposed model. In general, the time should be more than the

models with prerequisite settings. But it is better to show the difference, which will benefit the user to make the decision and trade-off.

Another thing I noticed the evaluation metrics were different for different tasks, and maybe it is because of the particularity of the different task. But I think it is better to use the uniform metrics to do the evaluation for different tasks or choose some similar tasks to compare with the uniform metrics.

7 Future Work

One thing for future work I think is to reduce the model time cost, for some tasks they care more about the time issues. Maybe we can make the model adapt from the proposed mix Gamma-negative binomial process to the basic process after the model get a reasonable topic number even it is not a best one, which can save some time in general. Besides, the MCMC method like Gibbs sampling-based inference is also slow, we can try to use other inference method in future.

Secondly, I think is to adapt the model for hierarchical multi-label problem, which labels are hierarchically structured like a tree or a graph. Maybe we need to re-design the Gamma Process structure. Besides, for text streaming we also need to figure out the time-evolving issues.

References

[1] J. Xuan, J. Lu, G. Zhang, R. Y. Da Xu, and X. Luo, "A bayesian nonparametric model for multi-label learning," *Machine Learning*, vol. 106, no. 11, pp. 1787–1815, 2017.