Literature Survey and Implementation: Bayesian non-exhaustive Online Learning
CS772A Project Proposal

Avikalp Kumar Gupta
Department of Computer Science & Engineering
IIT Kanpur
avikalpg@iitk.ac.in

Deepanshu Gupta
Department of Mathematics & Statistics
IIT Kanpur
dpanshu@iitk.ac.in

Jay Pandya
Department of Electrical Engineering
IIT Kanpur
jpandya@iitk.ac.in

1 Problem Statement

"Problem to be addressed in the project is of data inference. With data from recognized (i.e. target classes present in the training data) as well as unrecognized classes, we cluster streaming instances."

The motivation for solving this problem arises from, but is not limited to, the problem of detecting new pathogens in human body. As we know, micro-organisms have high adaptability and mutation rate, because of which new types of pathogens are found very frequently. Development of reagent-less techniques is hence very desirable for bio-security.

Similar techniques are also used for online document clustering, as suggested by [Zhang et al., 2004]. Article and Document tagging is a very actively researched problem. With every document having unspecified number of tags, and with every new document, emergence of a new tag is highly probable. Thus, we need a methodology that can adapt to emerging new tags and predict them at the same time.

2 Previous Work

Non-parametric Probabilistic Models are quite popular in Novelty Detection Literature. A lot of research has been done in the field, [Pimentel et al., 2014] have surveyed non-parametric techniques (especially Kernel density and Negative Selection based approaches) in novelty detection.

A good amount of work in Online document clustering has appeared in the literature. [Zhang et al., 2004] employed Dirichlet Process Prior to account for the growing number of clusters and utilized a Bayesian based approach to cluster the documents.

Significant prior work includes [Dundar et al., 2012], which proposes a new Sequence Importance Re-sampling (SIR) technique. They simply couple a Normal data model along with the Dirichlet Process Prior. This techniques evaluates the probability of assigning a new instance to an emerging class without the knowledge of the previous observed instances.

3 Proposal

The project will be a literature survey of applications of Dirichlet Process Prior for clustering, especially in online setting for non-exhaustive data. The project will also have a comparative study between performance of the proposed methodology in exhaustive vs non-exhaustive setting.

We follow the framework laid down by [Dundar et al., 2012] in which the authors infer the presence of a non-exhaustively defined set of classes in an on-line setting. The paper uses Dirichlet process prior (DPP) model over class distributions. DPP is very popular in the non-parametric Bayesian setting when discovering new classes (example: Zero Shot Learning).

References

[Dundar et al., 2012] Dundar, M., Akova, F., Qi, A., and Rajwa, B. (2012). Bayesian nonexhaustive learning for online discovery and modeling of emerging classes. *arXiv preprint arXiv:1206.4600*.

[Pimentel et al., 2014] Pimentel, M. A., Clifton, D. A., Clifton, L., and Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, 99:215–249.

[Zhang et al., 2004] Zhang, J., Ghahramani, Z., and Yang, Y. (2004). A probabilistic model for online document clustering with application to novelty detection. In *Advances in Neural Information Processing Systems*, pages 1617–1624.