A Visual Tool for Bayesian Data Analysis: The Impact of Smoothing on Naïve Bayes Text Classifiers

Giorgio Maria Di Nunzio
Dept. of Information Engineering
University of Padua
dinunzio@dei.unipd.it

Alessandro Sordoni
Dépt. d'Informatique et Recherche Opérationnelle
Université de Montréal
sordonia@iro.umontreal.ca

ABSTRACT

Naïve Bayes (NB) classifiers are simple probabilistic classifiers still widely used in supervised learning due to their tradeoff between efficient model training and good empirical results. One of the drawbacks of these classifiers is that in situations of data sparsity (i.e. when the size of training set is small) the maximum likelihood estimation of the probability of unseen features in these situations is equal to zero causing arithmetic anomalies. To prevent this undesirable behavior, a number of smoothing techniques have been proposed [4]. Among these, the Bayesian approach incorporates smoothing in terms of prior knowledge about the parameters of the model usually called *hyper-parameters*. Our research question is: can a visualization tool help researchers to quickly assess the goodness of the performance of NB classifiers by setting optimal smoothing parameters?

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Retrieval Models

General Terms: Theory, Algorithms, Experimentation

Keywords: Visual Analytics, Bayesian Inference, Smoothing, Text Classification

1. DEMONSTRATION

Our demo aims to record and visualize the effect of different combinations of choices of NB classifiers and smoothing approaches. Inspired by the work of [2, 3, 4], this tool addresses researchers who want to (i) compare different smoothing variants with a manual optimization of the parameters by means of the two-dimensional visual approach [1] and (ii) investigate the differences between Bayesian prior smoothing and the fixed coefficient interpolation method [4]. Moreover, the graphical visualization may be helpful for educational purposes or research projects that consider to employ NB classifiers.

Three NB classifiers — multivariate Bernoulli model, Multinomial model, Poisson model — can be selected and the relative priors' hyper-parameters (Beta, Dirichlet, and Gamma prior respectively) can be tuned to produce different smoothing effect on different standard test collection (Reuters, 20-Newsgroups, Oshumed). The visualization tool allows the user to (i) direct input the values of the hyper-parameters to compare standard smoothing variants, (ii) recall specific settings of actual systems implemented in literature and (iii) perform a visual search of best hyper parameter settings by analyzing their impact on the decision frontier on-the-fly. The visualization tool was designed following the *Model-View-Controller* pattern and was fully implemented in Java 1.6. This architecture is fully modular and allows for easy integration of new models and views. The source code can be found at `https://bitbucket.org/2dpm/2dpm`.

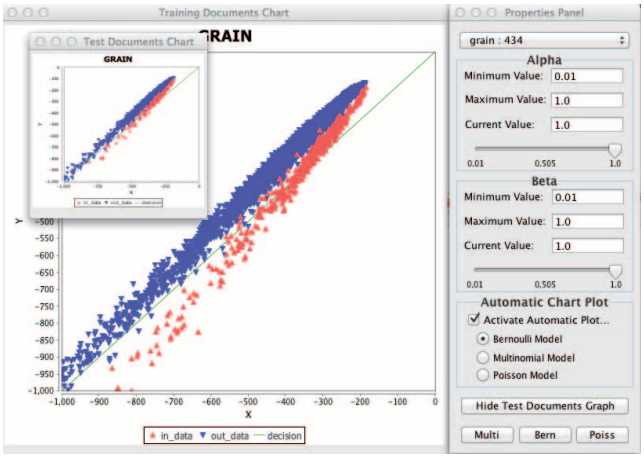

Figure 1: Two of the main panels are shown: the properties panel has been set to recall the multivariate Bernoulli model with Laplacian smoothing.

2. ACKNOWLEDGMENTS.

The authors would like to thank Prof. Jian-Yun Nie for the useful discussions. This work has been partially supported by the QONTEXT project under grant agreement N. 247590 (FP7/2007-2013).

3. REFERENCES

[1] G. Di Nunzio. Using scatterplots to understand and improve probabilistic models for text categorization and retrieval. *Int. J. Approx. Reasoning*, 50(7):945–956, 2009.

[2] S. Eyheramendy, D. D. Lewis, and D. Madigan. On the Naive Bayes Model for Text Categorization. In C. Bishop and B. Frey, editors, *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, 2003.

[3] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48, 1998.

[4] C. Zhai and J. D. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.