# Comparative Classifier Evaluation for Web-scale Taxonomies using Power Law

Rohit Babbar, Ioannis Partalas, Cornelia Metzig, Eric Gaussier, and Massih-reza Amini

Laboratoire d'Informatique de Grenoble, Université Joseph Fourier, Grenoble, France
{firstname.lastname}@imag.fr

**Abstract.** In the context of web-scale taxonomies such as Directory Mozilla(www.dmoz.org), previous works have shown the existence of power law distribution in the size of the categories for every level in the taxonomy. In this work, we analyse how such high-level semantics can be leveraged to evaluate accuracy of hierarchical classifiers which automatically assign the unseen documents to leaf-level categories. The proposed method offers computational advantages over $k$-fold cross-validation.

## 1 Introduction

The existence of power law for explaining the underlying phenomena has been observed in a variety of domains including sizes of cities [3], internet topologies [1]. In large scale textual hierarchies, the distribution of documents in nodes at individual levels, is also shown to exhibit similar behavior [2]. Directory Mozilla, a rooted tree taxonomy, lists over 5 million websites among 1 million categories and is maintained by close to 100,000 human editors. Hence, there is a need for automatic classification of unseen documents to the desired categories. In this work, we propose an approach which captures the high level semantics of such taxonomies to efficiently evaluate the performance of a hierarchical classifier.

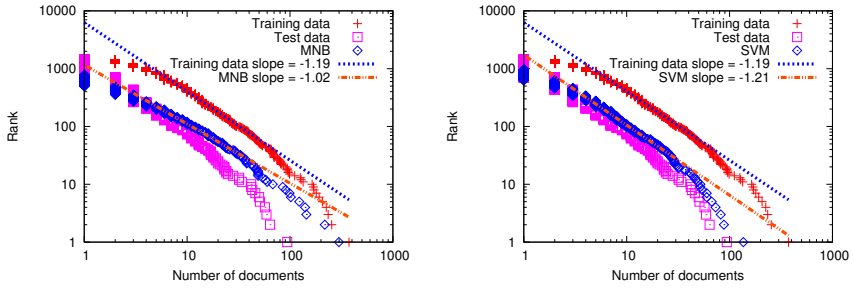
## 2 Proposed Approach

For the purpose of our work, we assume taxonomies in the form of rooted-tree in which the training documents are at leaves. The training documents for a classifier at a particular node in the taxonomy consist of all the documents of the leaves in the subtree rooted at that node. The number of training documents in the $j$-th ranked category (according to number of documents) for level $i$, denoted by $N_{ij}$, can be represented by [2]: $N_{ij} \approx N_{i1} j^{-\alpha_i}$, where $\alpha_i > 0$ is a level specific parameter. This distribution is shown in Figure 1, for level-5 of the DMOZ dataset extracted from the LSHTC-2011 challenge (http://lshtc.iit.demokritos.gr/).

The proposed strategy relies on the common assumption in machine learning that training and test data come from the same distribution. As shown in Figure 1, the curves for training and test data represented by plus and square signs receptively are parallel to each other. Using this intuition to compare the

relative accuracies of two hierarchical classifiers, the curve for the better classifier is likely to be more parallel to that of training data. Suppose $C^*$, represents the current best classifier, for a new classifier $C'$ to be better than $C^*$:
***Condition***: The power law exponent of $C'$ should be more closer to that of training data than the power law exponent of $C^*$.
The commonly used method of $k$-fold cross-validation, on other hand, is computationally expensive especially for web-scale data having tens of thousands of categories. Using a separate held-out set for the purpose of classifier evaluation leads to wasting training data which is scarce for rare categories as shown in [2].

***Experiments.*** We performed experiments with 50,538 training documents among 4,787 leaves in a tree of depth 6. Support Vector Machine (SVM) and Multinomial Naive Bayes (MNB) classifiers were trained to evaluate the performance on a test set of size 13,057. Figures 1(a) and 1(b) show the distribution of test documents among categories after classification by MNB and SVM respectively. Two conclusions are apparent by observing Figure-1:
1) The plot for SVM in Figure 1(b) is much more parallel to training data (and test data) than MNB, supporting the fact that SVM with accuracy of 52.3% is better than MNB whose accuracy is 36.5%.
2) The number of documents in the highest ranked category are 136 and 299 for SVM and MNB respectively, and the true value is 94.




(a) Power law plot for MNB (b) Power law plot for SVM

**Fig. 1.** Power Plots for MNB and SVM

## References


1. M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *SIGCOMM*.
2. T.-Y. Liu, Y. Yang, H. Wan, H.-J. Zeng, Z. Chen, and W.-Y. Ma. Support vector machines classification with a very large-scale taxonomy. *SIGKDD*, 2005.
3. M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 2005.