HIClass: Hyper Interactive text Classification by interactive supervision of document and term labels

Shantanu Godbole, Abhay Harpale, Sunita Sarawagi, and Soumen Chakrabarti

IIT Bombay
Powai, Mumbai, 400076, India
Contact: shantanu@it.iitb.ac.in

Abstract. We present the HIClass (Hyper Interactive text Classification) system, an interactive text classification system which combines the cognitive power of humans with the power of automated learners to make statistically sound classification decisions. HIClass is based on active learning principles and has aids for detailed analysis and fine tuning of text classifiers while exerting a low cognitive load on the user.

1 Introduction

Motivated by applications like spam filtering, e-mail routing, Web directory maintenance, and news filtering, text classification has been researched extensively in recent years [1–3]. Most text classification research assumes a simple bag-of-words model of features, a fixed set of labels, and the availability of a labeled corpus that is representative of the test corpus. Many of these assumptions do not hold in real-life.

Discrimination between labels can be difficult unless features are engineered and selected with extensive human knowledge. Often, there is no labeled corpus to start with, or the label set must evolve with the user's understanding. Projects reported routinely at the annual OTC workshops [4] describe applications in which automated, batch-mode techniques were unsatisfactory; substantial human involvement was required before a suitable feature set, label system, labeled corpus, rule base, and system accuracy were attained. Not all commercial systems use publicly known techniques, and few general principles can be derived from them.

There is scope for building learning tools which engage the user in an active dialog to acquire knowledge about features and labels. We present the HIClass system which provides a tight interaction loop for such an active dialog with the expert. *Active learning* has provided clear principles [5–7] and strategies for maximum payoffs from such a dialog. We extend active learning to include feature engineering and multi-labeled document labeling conversations. HIClass is an interactive multi-class multi-labeled text classification system that combines the cognitive power of humans with the power of automated learners to make statistically sound classification decisions (details appear in [8]).

2 The HIClass workbench for text classification

We present an overview of HIClass in Fig. 1. The lower layer shows the main data entities and processing units. There is a small labeled pool and a large unlabeled pool of documents. The system stores and accesses by name, multiple classifiers with their parameters, for comparative analysis and diagnostics. The upper layer shows main modes/menus of interaction with the system. We outline major components of the system next.

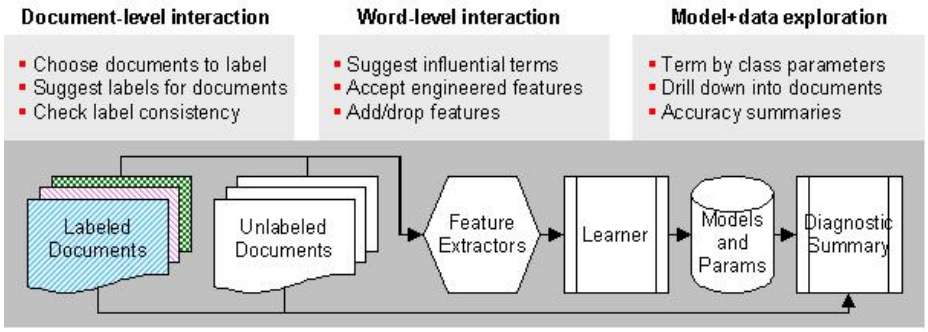

Fig. 1. The architecture of HIClass

Document and classification models: HIClass is designed using a flexible classification model template that (1) suits state-of-the-art automated learners and (2) can be easily interpreted and tuned by the user. HIClass uses **linear additive** classifier models like linear SVMs.

Exploration of data/models/performance summaries: HIClass allows the user to view the trained classifier scores, aggregate and drill-down statistics about terms, documents and classes, and different accuracy measures. An OLAP-like tool enables the human expert to fine tune individual features contributing to various classes with continuous feedback about resultant system accuracy.

Feature Engineering: Expert users, on inspection of class-term scores, will be able to propose modifications to the classifiers like adding, removing, or ignoring certain terms for certain classes. They can also provide input about stemming, aggregation, and combination of features.

Document labeling assistant: HIClass maximizes learning rate while minimizing user's cognitive load through various mechanisms: (1) a pool of most uncertain unlabeled documents is selected for user feedback, (2) bulk-labeling is facilitated by clustering documents, (3) the system ranks suggested labels, (4) the system checks labeling conflicts.

Term-level active learning: Initially when bootstrapping from a small corpus, HIClass directly asks users to specify well known trigger features as being positively/negatively associated with classes. This exerts a lower cognitive load on the user compared to reading full documents.

3 Description of the demonstration

Our demonstration will showcase the detailed working of all aspects of the HIClass system. HIClass consists of roughly 5000 lines of C++ code for the back-end which communicates through XML with 1000 lines of PHP scripts to manage browser-based front-end user interactions [8].

We will allow user interaction along three major modes. First, the user can either bootstrap the classifier by term-based active learning, or engage in traditional document-level labeling. Various labeling aids will minimize cognitive load on the user. The second major mode will be inspection of named learned classifiers in an OLAP-like interface for feature engineering. In the third exploratory mode, various aggregate statistics will draw the user's attention to areas which can benefit by more data and fine tuning. We will present extensive experimental results on benchmark text datasets highlighting the various aspects of the HIClass system.

References

1. T. Joachims. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of ECML 1998*.
2. K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification, 1999.
3. J. Zhang and Y. Yang. Robustness of regularized linear classification methods in text categorization. In *SIGIR 2003*.
4. 3rd workshop on Operational Text Classification OTC 2003. At SIGKDD-2003
5. D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. In *Advances in Neural Information Processing Systems*, 1995.
6. Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.
7. S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2:45–66, Nov. 2001.
8. S. Godbole, A. Harpale, S. Sarawagi, and S. Chakrabarti. Document classification through interactive supervision on both document and term labels In *Proceedings of PKDD 2004*.