In Proceeding of the 2005 Annual Symposium of the American Medical Informatics Association (AMIA 2005), October 2005, Washington, D.C. (poster)

"Bag of Words" is not enough for Strength of Evidence Classification

Jimmy Lin, Ph.D. and Dina Demner-Fushman, M.D., Ph.D.
University of Maryland, College Park

Abstract

Incorporation of evidence from clinical research requires critical appraisal of its quality. Information retrieval systems can facilitate physicians' judgments by automatically labeling retrieved citations with their strength of evidence categories. Preliminary results of such a text classification experiment involving MEDLINE® citations show that a "bag of words" approach is insufficient for accurate classification.

Introduction

One key step in the practice of evidence-based medicine (EBM) is incorporating best available research evidence into the clinical decision-making process [1], which involves appraisal of the information for validity and relevance. Guidelines for rating of the strength of evidence include three key elements: quality, quantity, and consistency [2]. Study type is one of the important factors contributing to the grading of research. For example, the Strength of Recommendation Taxonomy [3] considers randomized clinical trials, meta-analysis, and cohort studies of high quality as the highest grade evidence (level 1); case-control studies, case-series, and prospective studies with poor follow-up (level 2) are less valuable for evidence-based medicine.

Previously, McKeown et al. [4] have focused on categorizing articles according to whether or not they represent a clinical study, using terms from the full text and other features such as the article length and structure. Our study expands on this work in two significant ways: 1) Since the full texts of articles are not always available, we attempt to automatically classify citations using only abstract text; 2) Recognizing that a binary classification is not sufficiently fine-grained, we propose a three-way classification based on evidence grades (level 1, level 2, other). Both assumptions make this task more difficult, but more realistic because it fits directly into the practice of evidence-based medicine.

Methods and Results

We employed a standard supervised machine learning approach for our experiments. 525,938 MEDLINE records from April 2002 to April 2003 were used to train a Naïve Bayes classifier, using MeSH and Publication Type metadata as the ground truth labels. Each abstract was represented as a "bag of words", where each stemmed term represented a feature. We chose as features the top fifty most discriminating terms with respect to each class, as measured by information gain. Three experiments were conducted, one involving three-way classification, and two binary classifications: grade 1 *vs.* other and grade 1+2 *vs.* other. All experiments involved 10 fold cross-validation.

For three-way classification, we obtained an accuracy of 68%. For grade 1 *vs.* other, 90%, and for grade 1+2 *vs.* other. 73%. Our classifier performed worse than the (not-so-useful) baseline of simply guessing the most common label: the prior for evidence level 1 is 4.9%; level 2, 16.4%; and neither, 78.7%.

Conclusions

Two conclusions can be drawn from our preliminary study. Due to the rarity of "good" citations in a large representative sample of MEDLINE citations, automatic classification by evidence grades is a difficult problem, especially using only abstract text (4 to 7 percent improvement of full text-based over abstract-based classification was shown in [5]). As a point of comparison, McKeown et al. employed a selected subset of cardiology articles and obtained much higher classification accuracy. While a "bag of words" representation with Naïve Bayes is considered a strong baseline for text classification, it is insufficient for our task. These results point to the need for advanced natural language processing techniques, along the lines of [6].

References

[1] Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-based medicine: how to practice and teach EBM. Second edition. New York: Churchill Livingston; 2000.

[2] AHRQ. Systems to rate the strength of scientific evidence. Summary, evidence report/technology assessment: number 47. Publication no. 02-E015; March, 2002.

[3] Ebell MH, Siwek J, Weiss BD, Woolf SH, Susman JL, Ewigman B, Bowman M. Strength of Recommendation Taxonomy (SORT): A patient-centered approach to grading evidence in medical literature. J Fam Pract. 2004; 53(2):111-20.

[4] McKeown KR, Elhadad E, Hatzivassiloglou V. Leveraging a common representation for personalized search and summarization in a medical digital library. JCDL 2003.

[5] Gay CW, Aronson AR, Kayaalp M. Semi-automatic indexing for online biomedical journals. AMIA 2005 (to appear).

[6] Wilcox A and Hripcsak G. Medical text representation for inductive learning. AMIA 2000.