

Towards Automating the Initial Screening Phase of a Systematic Review

Tanja Bekhuis^a and Dina Demner-Fushman^b

^a Center for Dental Informatics, School of Dental Medicine, University of Pittsburgh, PA

^b Communications Engineering Branch, Lister Hill National Center for Biomedical Communications,
US National Library of Medicine, Bethesda, MD

Abstract

Systematic review authors synthesize research to guide clinicians in their practice of evidence-based medicine. Teammates independently identify provisionally eligible studies by reading the same set of hundreds and sometimes thousands of citations during an initial screening phase. We investigated whether supervised machine learning methods can potentially reduce their workload. We also extended earlier research by including observational studies of a rare condition. To build training and test sets, we used annotated citations from a search conducted for an in-progress Cochrane systematic review. We extracted features from titles, abstracts, and metadata, then trained, optimized, and tested several classifiers with respect to mean performance based on 10-fold cross-validations. In the training condition, the evolutionary support vector machine (EvoSVM) with an Epanechnikov or radial kernel is the best classifier: mean recall=100%; mean precision=48% and 41%, respectively. In the test condition, EvoSVM performance degrades: mean recall=77%, mean precision ranges from 26% to 37%. Because near-perfect recall is essential in this context, we conclude that supervised machine learning methods may be useful for reducing workload under certain conditions.

Keywords:

Artificial intelligence; machine learning; review literature as topic; review, systematic; study characteristics [publication type]; Cochrane Oral Health Group

Introduction

We conducted this study to demonstrate that supervised machine learning methods can potentially reduce the workload of systematic reviewers during the initial screening phase of citations. In this phase, teammates independently identify provisionally eligible studies by reading the same set of hundreds and sometimes thousands of titles and abstracts (TIABS). This bottleneck slows the production of quality systematic reviews meant to synthesize research to guide clinicians in their practice of evidence-based medicine.

Additionally, we extended the work of Aphinyanaphongs *et al.* [1], Cohen *et al.* [2], and Kilicoglu *et al.* [3] who sought to find

rigorous clinical research using supervised machine learning methods. Based on the work of Haynes and colleagues (e.g., see [4]), rigor was presumed if trials comparing treatments were randomized and controlled.

To classify studies with respect to rigor or quality, each research group constructed a reference collection or ‘gold standard’ of positive cases. Aphinyanaphongs and colleagues [1] used MEDLINE records for articles abstracted by the *ACP Journal Club*, which is a respected meta-journal that abstracts or cites evidence-based research in internal medicine for clinicians. Cohen *et al.* [2] used citations for randomized controlled trials (RCTs) included in 15 systematic reviews of drug therapies conducted by an Evidence-based Practice Center (EPC) funded by the US Agency for Healthcare Research and Quality. (The EPC files are publicly available at <http://medir.ohsu.edu/~cohenaa/systematic-drug-class-review-data.html>.) Kilicoglu *et al.* [3] used a large subset of manually annotated citations for documents that were used to develop the clinical query filters in PubMed [4]. They selected rigorous studies relevant to human healthcare with a treatment or prevention focus as a gold standard.

Because randomized and quasi-randomized controlled trials (RCTs) tend to be less biased relative to nonrandomized and observational studies, review authors prefer to include RCTs and quasi-RCTs in their systematic reviews. However, it is sometimes necessary to include studies with weaker designs when RCTs are unlikely or unethical. For example, nonrandomized and observational studies are common for studies of: exposure to environmental hazards; invasive surgery compared to no surgery; risk factors for patients with chronic conditions; outcomes associated with patient-selected devices or over-the-counter drugs; diagnostic accuracy; and rare disorders. Thus, to meet the need for synthesized evidence for these kinds of questions, classification methods for studies with weaker designs should be developed along with those for RCTs.

The challenges are significant. Consider, for example, that abstracts with potentially informative words and phrases regarding trial or study design were unavailable in MEDLINE for most articles published before 1976. Moreover, few terms

for designs were available in the MeSH Thesaurus before the 1990s [5 (p. 131)]. Since then, terms have been added or modified to index designs, including weaker ones. For example, the term ‘study characteristics [publication type]’ includes narrower terms for ‘case reports,’ ‘comparative study,’ and ‘evaluation studies.’ Nevertheless, according to the Cochrane Non-Randomised Studies Methods Group, (1) authors of primary studies inconsistently describe the designs of their studies; (2) bibliographic databases do not reliably index designs; and (3) good filters for nonrandomized or observational studies do not yet exist [6]. In fact, when the latter are eligible for inclusion in a review, authors are enjoined to *not* include design terms in their search filters unless the retrieval set is so large that the review becomes impractical. Thus, the initial screening phase is typically labor intensive when both randomized trials and nonrandomized studies are eligible.

Methods

We used a recently approved search strategy for a Cochrane systematic review about ameloblastomas, which are rare odontogenic tumors of the jaws [7]. This strategy combines a topic filter with the Cochrane highly sensitive filter for identifying randomized controlled trials (see Box 6.4.c in [5]), and a modified SIGN filter for observational studies [8]. (Without terms for designs, the size of the initial retrieval set would have forestalled the review.) The combined filter is designed to find studies that compare surgical resection to any other treatment of ameloblastomas. The editors of the Cochrane Oral Health Group acknowledged the probable low incidence of ameloblastomas and therefore approved inclusion of case-control and patient registry studies. Because the primary outcome is recurrence of the tumor, Bekhuis and colleagues modified the SIGN filter by excluding cross-sectional studies and by including terms for registry studies [7].

The Cochrane Oral Health Group Trials Search Coordinator conducted the search, which yielded 1774 citations from four databases: MEDLINE, EMBASE, the Cochrane Central Register of Controlled Trials (CENTRAL), and the Cochrane Oral Health Group Trials Register. We also retrieved 41 citations from two systematic reviews [9, 10]. After de-duplication, the total number of citations was 1814. We sorted the corpus by publication date in descending order. Even though indexing may be inadequate with respect to design, the sorting reflects our belief that observational studies published after 2007 may be better described in titles and abstracts. This is partly because of the increasing adoption of the STROBE statement for reporting observational studies [11] by biomedical journals, including *Annals of Internal Medicine*, *Lancet*, and *PLoS*, among others. (See a list of journals at <http://www.strobe-statement.org>.) In the STROBE checklist, one of several recommendations for writing a good report states that authors should “indicate the study’s design with a

commonly used term in the title or the abstract.” The checklist is available at <http://www.strobe-statement.org/index.php?id=checklists>.

We built training and test sets by selecting the most recent citations from the initial retrieval set and then proportionately distributed citations from the systematic reviews. Citations in the test set (n=100) and training set (n=300) were labeled with respect to eligibility status in accordance with the consensus decisions of the Cochrane review team [7]. Thus, citations pointing to provisionally eligible studies were labeled as ‘include’ and those pointing to ineligible studies as ‘exclude.’ Thirteen percent of studies were provisionally eligible in both the training and test sets.

We used EndNote to manage citations, to record eligibility decisions of the review team, and to export a text file of 400 citations which was then ‘chunked’ into separate files (one per citation) using Perl. We used RapidMiner [12], a software package for machine learning and data mining, to which we added a plug-in to process text (available at <http://wvtool.sourceforge.net>).

Features were extracted from TIABS and metadata using a bag-of-words approach. Pre-processing text involved string tokenizing, converting to lower case, filtering out Medline [13] or English stop words, filtering out tokens with length less than 3, and Porter stemming. Feature vectors were weighted with term frequencies (TF) or the product of TF and inverse document frequencies (TFIDF); vectors were pruned of terms that occurred in at most 3 citations. Features were selected for information gain.

Broadly, we followed the following steps: (1) We trained several classifiers using processed feature sets; (2) compared mean performance of classifiers based on 10-fold cross-validations, where performance measures were mean recall, mean precision, and the harmonic mean of equally-weighted precision and recall (F_1); (3) used grid optimization to find the kernel type that minimized absolute error for the evolutionary support vector machine (EvoSVM) classifier [14]; (4) investigated the impact of training set size on performance; and (5) compared the performance of EvoSVM configurations on the held-out test set.

Results

In early analyses, naïve Bayes and support vector machines (SVMs)—distinct from EvoSVMs—failed as classifiers, even though many researchers have used these algorithms to successfully classify documents [15]. Instead, we compared the following RapidMiner classifiers: DecisionTree, EvoSVM, and weightily averaged one-dependence estimator (WAODE) [16], focusing on EvoSVM in later analyses. To train the WAODE classifier, we first discretized features using the

minimal entropy partitioning operator. Selected training results are presented in Table 1.

Analyses not presented compared the performance of each classifier by varying the weights (TF vs TFIDF) for the feature vectors. With the exception of the DecisionTree classifier, performance was better when using TFIDF weights. English stop words instead of MEDLINE stop words were used when pre-processing text for all classifiers because recall was higher when using the former in early analyses.

Table 1–Mean training performance of selected classifiers over 10-fold cross-validations

Classifier	Performance		
	Mean Recall (%)	Mean Precision (%)	F ₁
DecisionTree (TF)			
MEDLINE Stop words	25.83	42.83	0.305
English Stop words	30.83	45.83	0.355
EvoSVM (TFIDF)			
Radial	100.00	41.47	0.578
Polynomial Degree 3	66.67	72.83	0.660
Polynomial Degree 4	65.83	73.50	0.676
Epanechnikov Degree 3	95.00	60.20	0.714
Epanechnikov Degree 4	100.00	48.29	0.648
WAODE (TFIDF)	65.83	72.33	0.677

In the training condition, recall is perfect for the EvoSVM classifier with a radial or Epanechnikov (degree 4) kernel, although precision is modest. F₁ is highest for the EvoSVM classifier with an Epanechnikov (degree 3) kernel (see Table 1).

Four EvoSVM kernel types (radial, Epanechnikov, Gaussian-combination, and multiquadric) were compared using a grid parameter optimization algorithm with 3 iterations over 10-fold cross-validations. The EvoSVM classifier with a radial kernel outperforms other configurations when considering absolute error, mean recall, and mean precision (see Table 2).

Table 2–Grid parameter optimization of EvoSVM kernel type (Complexity=1; sigma 1=10; TFIDF)

EvoSVM Kernel	Performance		
	Absolute Error	Mean Recall (%)	Mean Precision (%)
Radial	0.253	92.3	75.0
Epanechnikov	0.265	90.4	72.2
Gaussian-Combination	0.535	28.8	39.5
Multiquadric	0.393	50.0	43.3

When we trained the EvoSVM classifier with an Epanechnikov kernel (degree 4) on 150 citations instead of 300, mean recall degraded considerably, but precision and F₁ improved: mean recall=70.00%, mean precision=76.47% and F₁=0.72.

To understand the impact of training set size, we compared the corresponding feature set size for n_{train}=300, 250, 200, 150, and 100 (see Figure 1). We used stratified sampling to preserve the proportion of provisionally eligible studies in each sample.

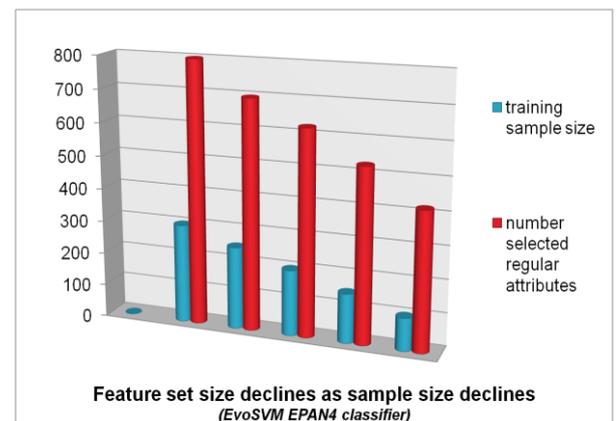


Figure 1–Feature set size is related to the number of citations in the training set.

We further investigated the relationship between performance of the EvoSVM classifier (Epanechnikov kernel, c=1, sigma=10, TFIDF) and training set size. Training sets were again stratified. Mean recall degrades as the size of the training set decreases, dropping markedly when n_{train} =100; precision peaks when n_{train} =200 (see Figure 2).

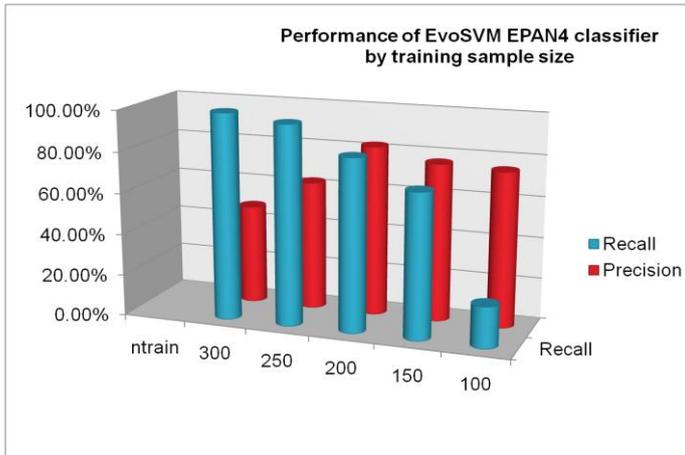


Figure 2—Performance of the EvoSVM classifier is related to the number of citations in the training set.

When we tested the EvoSVM classifier on the held-out test set of citations, performance degraded. Mean recall is equivalent for three configurations (77%), mean precision ranges from 26% to 37%, and F_1 from 0.39 to 0.50 (see Table 3).

Table 3—Performance of the EvoSVM classifier on the held-out test set of citations

EvoSVM Kernel	Performance		
	Mean Recall (%)	Mean Precision (%)	F_1
Radial	76.92	26.32	0.392
Epanechnikov Degree 3	76.92	37.04	0.500
Epanechnikov Degree 4	76.92	29.41	0.426

Note. Further analyses were conducted and the results are available upon request.

Discussion

It is important to realize that recall must be optimal for any machine learning approach meant to aid systematic review authors. For example, the Cochrane Collaboration strongly recommends broad and sensitive search strategies with high recall so that relevant research is not overlooked. In addition, review authors must make good-faith efforts to locate research missed by electronic searches. Thus, they handsearch journals, scan reference lists, contact subject experts, and more. This is why attaining very high recall is our primary goal. Boosting precision is a secondary goal even though modest precision is not as problematic as one might think when the percentage of provisionally eligible studies is relatively low. However, it may

be a problem when the percentage is relatively high. Consider the following scenarios.

1. Assume that 2000 citations are retrieved, 10% point to provisionally eligible studies, recall is perfect (100%), and precision is modest (50%). The classifier will correctly include 200 citations and incorrectly include another 200. The second review author of a two-member review team has to read 400 TIABS instead of 2000, which reduces her workload by 80%.
2. Same assumptions as before, except that 40% point to provisionally eligible studies. The classifier will correctly include 800 citations and incorrectly include another 800. The second review author has to read 1600 citations, which reduces her workload by 20%.

Nevertheless, the absolute reduction of workload is probably more important to a human than the percent reduction. Consider, for example, that in the second scenario just posed, the review author is spared reading 400 TIABS even though her workload is reduced by just 20%.

Classifiers

In early analyses, the failure of naïve Bayes and support vector machine (SVM) classifiers—distinct from EvoSVMs—may have been due to violations of statistical assumptions. For example, naïve Bayes assumes independence of features and positional independence, and SVM assumes linearly separable classes. Because WAODE [16] and EvoSVM [14] classifiers relax these assumptions, they are more appropriate for these data. (The WAODE classifier differentially weights tree-augmented naïve Bayes models according to how informative each attribute is when set as the root of a tree. The EvoSVM classifier finds an optimal nonlinear hyperplane to classify data that are not linearly separable.)

Although we attained perfect recall with the EvoSVM classifier in the training condition and very high recall for EvoSVM with optimization, over fitting is still a concern. This concern was borne out in the held-out test condition when recall degraded. Nevertheless, the results are promising and suggest that EvoSVM with a radial or Epanechnikov kernel may be an appropriate classifier when observational studies are eligible for inclusion in a systematic review.

Limitations

This study has serious limitations. First, the extracted features may not have been representative of the domain because of the small size of the training set. This could account for the degradation of performance on the held-out test set. In the future, more than 1800 labeled citations from the initial screening phase of a Cochrane review [7] will be available. We expect that performance will improve when the classifiers are trained on a much larger set of citations than was the case for this study. Second, the bag-of-words approach—although affording an appropriate baseline—ignores important phrases, such as *case report*, *case series*, *literature review*, and *ameloblastomas of the jaws*. In the future, we will explore various feature sets to improve classification in the held-out testing phase. This will entail annotating citations for relevant terms and phrases, including design features and possibly

affiliation and journal. Third, we know that stacking (a method of weighting several classifiers) is a promising approach [3, 17]. However, stacking probably works best with diverse feature sets and is therefore a method more appropriate for a larger study. Finally, future evaluation of classifier performance needs to be statistically rigorous.

Conclusion

The evidence suggests that supervised machine learning methods can potentially reduce the workload of systematic review authors during the initial screening phase when (1) observational studies of treatments for a rare condition are eligible for inclusion in the review, (2) the proportion of provisionally eligible studies is relatively small, and (3) the number of citations is large enough to capture representative features.

Acknowledgments

This research was supported, in part, by the US National Library of Medicine (NLM) Research Participation Program, sponsored by NLM and administered by the Oak Ridge Institute for Science and Education; and by the NLM/NIDCR Pittsburgh Biomedical Informatics Training Program 5 T15 LM/DE07059-22. We would like to thank Mr. Halil Kilicoglu and Mr. Matthew Simpson for technical support; Drs. Thankam Thyvalikakath and Richard Oliver for help with labeling citations; and Ms. Anne Littlewood, Cochrane Oral Health Group Trials Search Coordinator, for conducting the search for citations.

References

- [1] Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis C. Text categorization models for high-quality article retrieval in internal medicine. *J Am Med Inform Assoc.* 2005;2:207-216.
- [2] Cohen AM, Hersh WR, Peterson K, Yen P-Y. Reducing workload in systematic review preparation using automated citation classification. *J Am Med Inform Assoc.* 2006;13:206-219.
- [3] Kilicoglu H, Demner-Fushman D, Rindfleisch TC, Wilczynski NL, Haynes RB. Towards automatic recognition of scientifically rigorous clinical research evidence. *J Am Med Inform Assoc.* 2009;16(1):25-31.
- [4] Wilczynski NL, Morgan D, Haynes RB, Hedges Team. An overview of the design and methods for retrieving high-quality studies for clinical care. *BMC Med Inform Decis Mak.* 2005;5(5):20.
- [5] Lefebvre C, Manheimer E, Glanville J. Chapter 6: Searching for Studies. In: Higgins JPT, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions.* Chichester (UK): Wiley, 2008.
- [6] Reeves BC, Deeks JJ, Higgins JPT, Wells GA. Chapter 13: Including non-randomized studies. In: Higgins JPT and Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions.* Chichester (UK): Wiley, 2008.
- [7] Bekhuis T, Thyvalikakath T, Oliver R. Interventions for treating ameloblastomas of the jaws [protocol in press]. *Cochrane Database Syst Rev.*
- [8] Scottish Intercollegiate Guidelines Network [SIGN]. Search Filters: Observational Studies (MEDLINE). <http://www.sign.ac.uk/methodology/filters.html>. Accessed 16 June 2008.
- [9] Lau SL, Samman N. Recurrence related to treatment modalities of unicystic ameloblastoma: a systematic review. *Int J Oral Maxillofac Surg.* 2006;35:681-690.
- [10] Pogrel MA, Montes DM. Is there a role for enucleation in the management of ameloblastoma? *Int J Oral Maxillofac Surg.* 2009. Doi:10.1016/j.ijom.2009.02.018.
- [11] von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The strengthening [of] the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *J Clin Epidemiol.* 2008;61:344-349.
- [12] Mierswa I, Wurst M, Klinkenberg R, Scholz M, Euler T. YALE (now RapidMiner): rapid prototyping for complex data mining tasks. In: *Proc ACM SIGKDD Int Conf on Knowl Discov Data Mining,* 2006.
- [13] Princeton University Fine Hall Biology Library. MEDLINE Data-base: Stop Words, 10/31/08. [File from Halil Kilicoglu.]
- [14] Mierswa I. Evolutionary learning with kernels: a generic solution for large margin problems. In: *Proc. of the Genetic and Evolutionary Computation Conference (GECCO),* 2006.
- [15] Sebastiani F. Machine learning in automated text categorization. *ACM Computing Surveys.* 2002;34(1):1-47
- [16] Jiang L, Zhang H. Weightily averaged one-dependence estimators. In: *Proc 9th Biennial Pacific Rim Int Conf Artificial Intelligence, PRICAI,* 2006; pp. 970-974.
- [17] Ting KM, Witten IH. Issues in stacked generalization. *J Artif Intell Res,* 1999;10:271-289.

Address for correspondence

Tanja Bekhuis, Center for Dental Informatics, 382 Salk Hall, 3501 Terrace Street, Pittsburgh, PA, 15261 US