LETTERS

Performance of support-vector-machine-based classification on 15 systematic review topics evaluated with the WSS@95 measure

In the July 2010 issue of JAMIA, Matwin *et al* published an article entitled 'A new algorithm for reducing the workload of experts in performing systematic reviews.'[1] Briefly, the work proposes a factorized variant of the complement Naïve Bayes classifier as an improvement, using weight engineering on the features (FCNB/WE). The prior work of Cohen *et al* in this area is cited, and the data set made public along with this prior work is used for the evaluation.[2]

The Matwin *et al* article compares the authors' proposed system against the early Cohen *et al* published voting perceptron (VP) classifier results, using the 'work saved over sampling at 95% recall' (WSS@95) measure proposed in that paper. However, the article notes that WSS@95 figures were not published in Cohen's later work based on the support vector machine (SVM) classifier,[3 4] and states that these figures were not available for comparison.

Our team has not published the WSS@95 figures for the SVM classifier previously because our current investigations focus on using classification for a number of different use cases within the systematic review process. Each of these use cases, and potentially each review team and user, may prefer different recall-precision trade offs, and therefore different classification thresholds. Furthermore, this may be affected by the size of the literature base in a given review domain and/or the prevalence of articles meeting the inclusion criteria for that review. Therefore, we have chosen to use the area under the receiver-operating curve (AUC) as our measure of classifier accuracy across the range of possible operating points, instead of the single point measure WSS@95. We are currently conducting a study of systematic reviewer preferences for classification performance within these use cases to better understand how to optimize for these trade-offs.

Although not previously published, in fact, we did collect WSS@95 performance figures for the 15 systematic review topic data set used by Matwin *et al*. The system used is identical to that published in our prior work.[4] These results were obtained using five repetitions of twofold crossvalidation, as in Matwin as well as our prior work, and are shown in table 1. As can be seen in the table, the performance of our previously published SVM-based system is generally superior to that of Matwin, outperforming the FCNB/WE system for 12 of 15 topics on WSS@95. On average, our previously published system increases WSS@95 by 0.073 over FCNB/WE. This represents a mean improvement of almost 22%.

Table 1 WSS@95 Comparison of Cohen 2008[4] and Matwin 2010[1] Systems across 15 systematic review topics using 5×2 cross-validation

Topic	Cohen 2008[4] support vector machine	Matwin 2010[1] FNCB/WE	Delta
ACE inhibitors	0.733	0.523	0.210
Attention deficit hyperactivity disorder	0.526	0.622	−0.096
Antihistamines	0.236	0.149	0.087
Atypical antipsychotics	0.170	0.206	−0.036
Beta blockers	0.465	0.367	0.098
Calcium-channel blockers	0.430	0.234	0.196
Estrogens	0.414	0.375	0.039
Nonsteroidal anti-inflammatory drugs	0.672	0.528	0.144
Opioids	0.364	0.554	−0.190
Oral hypoglycemics	0.136	0.085	0.051
Proton pump inhibitors	0.328	0.229	0.099
Skeletal muscle relaxants	0.374	0.265	0.109
Statins	0.491	0.315	0.176
Triptans	0.346	0.274	0.072
Urinary incontinence	0.432	0.296	0.136
Mean	0.408	0.335	0.073

FNCB/WE, Matwin *et al* 2010[1] factorized variant of the complement Naïve Bayes classifier as an improvement, using weight engineering on the features.

It is interesting that the SVM approach is inferior to our prior VP results for the attention deficit hyperactivity disorder (ADHD) topic, and that FCNB/WE is superior to both SVM and VP for the opioids topic, especially given that the SVM AUC measure is about 0.90 for both of these topics. Both the ADHD and opioids topics have very low article inclusion rates (2.4% and 0.8% respectively) and a relatively small number of positive samples (20 and 15 respectively). Clearly, there is an opportunity for future research in enhancing classifier performance at the very high recall end of the receiver-operating curve, especially with very small numbers of positive samples in the training data.

We are pleased to see other researchers investigating the important area of improving the efficiency of the systematic review process. We encourage Matwin *et al* to continue to improve their system, as we are, and to push the field forward, making these and other potentially effective tools available to support the systematic review process and evidence-based medicine.

Aaron M Cohen

Correspondence to Dr Aaron M Cohen, Department of Medical Informatics and Clinical Epidemiology, School of Medicine, Oregon Health & Science University, 3181 SW Sam Jackson Park Road, Mail Code: BICC, Portland, OR 97239-3098, USA; cohenaa@ohsu.edu

Funding National Institutes of Health; National Library of Medicine.

Provenance and peer review Not commissioned; not externally peer reviewed.

Received 7 September 2010
Accepted 5 November 2010

J Am Med Inform Assoc 2011;**18**:104.
doi:10.1136/jamia.2010.008177

REFERENCES

1. **Matwin S,** Kouznetsov A, Inkpen D, *et al*. A new algorithm for reducing the workload of experts in performing systematic reviews. *J Am Med Inform Assoc* 2010;**17**:446–53.
2. **Cohen AM,** Hersh WR, Peterson K, *et al*. Reducing workload in systematic review preparation using automated citation classification. *J Am Med Inform Assoc* 2006;**13**:206–19.
3. **Cohen AM,** Ambert K, McDonagh M. Cross-topic learning for work prioritization in systematic review creation and update. *J Am Med Inform Assoc* 2009;**16**:690–704.
4. **Cohen A.** Optimizing feature representation for automated systematic review work prioritization. *AMIA Annu Symp Proc* 2008:121–5.

Performance of SVM and Bayesian classifiers on the systematic review classification task

We are grateful to Professor Cohen for his letter and clarification of the support vector machine algorithm (SVM) results (in press). We agree that the results he supplies fill a gap in our paper. We could not have performed this comparison in our paper, as the first version was written prior to the publication of his own article,[1] which in any case, as Dr Cohen points out, did not include the SVM results in terms of within-groups sum of squares (WSS).

We would like, however, to be cautious about the broader conclusion from the results in table 1 of Dr Cohen's letter. While they are indeed superior to our factorized version of the complement naïve Bayes (FCNB) approach, they do not necessarily indicate the general superiority of the SVM