Classifying Biomedical Abstracts Using Committees of Classifiers and Collective Ranking Techniques

Alexandre Kouznetsov[1], Stan Matwin[1], Diana Inkpen[1], Amir H. Razavi[1], Oana Frunza[1], Morvarid Sehatkar[1], Leanne Seaward[1], Peter O'Blenis[2]
[1]School of Information Technology and Engineering (SITE), University of Ottawa
[2]TrialStat Corporation, 1101 Prince of Wales Drive, Ottawa, ON, CA, K2C 3W7
akouz086@uottawa.ca,{stan, diana, araza082, ofrunza, mseha092}@site.uottawa.ca,
lspra072@uottawa.ca, poblenis@trialstat.com

Abstract. The purpose of this work is to reduce the workload of human experts in building systematic reviews from published articles, used in evidence-based medicine. We propose to use a committee of classifiers to rank biomedical abstracts based on the predicted relevance to the topic under review. In our approach, we identify two subsets of abstracts: one that represents the top, and another that represents the bottom of the ranked list. These subsets, identified using machine learning (ML) techniques, are considered zones where abstracts are labeled with high confidence as relevant or irrelevant to the topic of the review. Early experiments with this approach using different classifiers and different representation techniques show significant workload reduction.

Keywords: Machine Learning, Automatic Text Classification, Systematic Reviews, Ranking Algorithms

1 Introduction

Evidence-based medicine (EBM) is an approach to medical research and practice that attempts to provide better care with better outcomes by basing clinical decisions on solid scientific evidence [1]. Systematic Reviews (SR) are one of the main tools of EBM. Building SRs is a process of reviewing literature on a specific topic with the goal of distilling a targeted subset of data. Usually, the reviewed data includes titles and abstracts of biomedical articles that could be relevant to the topic. SR can be seen as a text classification problem with two classes: a positive class containing articles relevant to the topic of review and a negative class for articles that are not relevant.
In this paper we propose an algorithm to reduce the workload of building SRs while maintaining the required performance of the existing manual workflow. The number of articles classified by the ML algorithm with high confidence can be considered a measure of workload reduction.

2. Ranking Method

Ranking Algorithm. The proposed approach is based on using committees of classification algorithms to rank instances based on their relevance to the topic of review. We have implemented a two-step ranking algorithm. While the first step,

called local ranking, is used to rank instances based on a single classifier output, the second step, named collective ranking, integrates the local ranking results of individual classifiers and sorts instances based on all local ranking results.

The local ranking process is a simple mapping:

$$R_j(w_{ij}^{+}, w_{ij}^{-}) \rightarrow s_{ij} \quad (1)$$

where R_j is the local ranking function for classifier *j*; w_{ij}^{+} and w_{ij}^{-} are decision weights for the positive and the negative class assigned by classifier *j* to instance *i*; s_{ij} is the local ranking score for instance *i* based on *c*lassifier's *j* output. Depending on what the classifier *j* is using as weights, s_{ij} are calculated as the ratio or normalized difference of the weights.

All instances to be classified (test set instances) are sorted based on the local ranking scores s_{ij}. A sorted list of instances is built for each classifier *j*. As a result, each instance *i* is assigned a local rank l_{ij} that is the position (the rank) of the current instance in the ordered list of instances with respect to the current classifier *j*:

$$s_{ij} \rightarrow l_{ij}, \qquad l_{ij} \in \{1, 2, \ldots, N\} \quad (2)$$

where *N* is the total number of instances to be classified.

In the second step, the collective ranking score g_i is calculated for each instance *i* over all the applied classifiers, as follows:

$$g_i = \sum_j (N - l_{ij} + 1) \quad (3)$$

All instances to be classified are in the end based on the collective ranking scores. The collective ordered list of instances is a result of this sorting. Finally, we get the collective rank r_i for each instance as the number associated with that instance in the collective ordered list (the position in the list):

$$g_i \rightarrow r_i, \qquad r_i \in \{1, 2, \ldots, N\} \quad (4)$$

An instance with a higher collective rank is more relevant to the topic under review than another instance with a lower collective rank.

Classification rule for the committee of classifiers. The classification decision of the committee is based on the collective ordered list of instances. The key point is to establish two thresholds:

T - top threshold (number of instances to be classified as positive);

B - bottom threshold (number of instances to be classified as negative).

We propose a special Machine Learning (ML) technique to determine *T* and *B* for new test data by applying our classifiers on the labeled data (the training set). Since human experts have assigned the labels for all training set instances, top and bottom thresholds on the training set could be created with respect to the required level of prediction confidence (which in our case is the average recall and precision level of individual human experts). As top and bottom thresholds for the training set are assigned, we simply project them on the test set, while adjusting them to the new

distribution of the data, the proportions of the size of the prediction zones and gray zone are maintained. After the thresholds are determined, the committee classification rule is as follows:

$$\begin{aligned} &(r_i \le T) => i \in Z^+, \quad c_i = relevant \\ &(r_i > (N - B)) => i \in Z^-, \quad c_i = irrelevant \\ &(T < r_i \le (N - B)) => i \in Z^N \end{aligned} \tag{5}$$

where c_i is final class prediction on instance *i*; r_i represents the collective rank of the instance *i* , *N* is a number of instances in the test set; Z^+ is the positive prediction zone the subset of the test set including all instances predicted to be positive with respect to required level of prediction confidence; Z^- is the negative prediction zone, the subset of the test set that consists of all instances predicted to be negative with respect to the required level of prediction confidence. The prediction zone is built as the union of Z^+ and Z^-. Test set instances that do not belong to the prediction zone belong to what we call the gray zone Z^N.

3. Experiments

The work presented here was done on a SR data set provided by TrialStat Corporation [2]. The source data includes 23334 medical articles pre-selected for the review. While 19637 articles have title and abstract, 3697 articles have only the title. The data set has an imbalance rate (the ratio of positive class to the entire data set) of 8.94%.

A stratified repeated random sampling scheme was applied to validate the experimental results. The data was randomly split into a training set and a test set five times. On each split, the training set included 7000 articles (~30%) , while the test set included 16334 articles (~70%) The results from each split were then averaged.

We applied two data representation schemes to build document-term matrices: Bag-of-words (BOW) and second order co-occurrence representation [3]. CHI2 feature selection was applied to exclude terms with low discriminative power.
In order to build the committee, we used the following algorithms[1]: (1) Complement Naïve Bayes [4]; (2) Discriminative Multinomial Naïve Bayes[5]; (3) Alternating Decision Tree [6]; (4) AdaBoost (Logistic Regression)[7]; (5)AdaBoost (j48)[7].

4. Results

By using the above described method to derive the test set thresholds from the training set, the top threshold is set to 700 and the bottom threshold is set to 8000. Therefore, the prediction zone consists of 8700 articles (700 top-zone articles and 8000 bottom-zone articles) that represent 37.3% of the whole corpus. At the same time, the gray zone includes 7634 articles (32.7% of the corpus). Table 1 presents the recall and precision results calculated for the positive class. (Only prediction zone

[1] We tried a wide set of algorithms with good track record in text classification tasks , according to the literature. We selected the 5 which had the best performance on our data.

articles are taken into account.) Table 1 also includes the average recall and precision results for human expert predictions[2], observed SR data from the TrialStat Inc.

The proposed approach includes two levels of ensembles: the committee of classifiers and an ensemble of data representation techniques. We observed that using the ensemble approach has brought significant impact on performance improving (possible because it removes the variance and the mistakes of individual algorithms).

Table 1. Performance Evaluation

Performance measure	Machine Learning results on the Prediction Zone	Average Human Reviewer's results
Recall on the positive class	91.6%	90-95%
Precision on the positive class	84.3%	80-85%

5. Conclusions

The experiments show that a committee of ML classifiers can rank biomedical research abstracts with a confidence level similar to human experts. The abstracts selected with our ranking method are classified by the ML technique with a recall value of 91.6% and a precision value of 84.3% for the class of interest. The human workload reduction that we achieved in our experiments is 37.3% over the whole data.

Acknowledgments. This work is funded in part by the Ontario Centres of Excellence and the Natural Sciences and Engineering Research Council of Canada.

References

1. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *BMJ* **312** (7023): 71-2. (1996)
2. TrialStat corporation web resources. Available at: http://www.trialstat.com/
3. Pedersen T., Kulkarni A., Angheluta R., Kozareva Z., Solorio T. An Unsupervised Language Independent Method of Name Discrimination Using Second Order Co-occurrence Features. Proceedings of the 7th International Conference on Intelligent Text Processing and Computational Linguistics, V. TBD of Lecture Notes in Computer Science, Springer, Mexico. (2006)
4. Rennie, J., Shih, L., Teevan, J., Karger, D., Tackling the poor assumptions of naive bayes text classifiers. In: ICML-2003, Washington DC. (2003)
5. Jiang Su, Harry Zhang, Charles X. Ling, Stan Matwin. Discriminative Parameter Learning for Bayesian Networks. In: ICML 2008. (2008)
6. Freund, Y., Mason, L.: The alternating decision tree learning algorithm. In: Proceeding of the 16 International Conference on ML, Slovenia, 124-133. (1999)
7. Freund Y, Schapire R: Experiments with a new boosting algorithm. In: Thirteenth International Conference on ML, San Francisco, 148-156. (1996)

[2] Experts are considered working individually. A few reviewers usually review each article. We partially replace one expert with a ML algorithm.