Using Classifier Performance Visualization to Improve Collective Ranking Techniques for Biomedical Abstracts Classification

Alexandre Kouznetsov[1] and Nathalie Japkowicz[2]

[1] Department of Computer Science and Applied Statistics,
University of New Brunswick Saint John
[2] School of Information Technology and Engineering, University of Ottawa

Abstract. The purpose of this work is to improve on the selection of algorithms for classifier committees applied to reducing the workload of human experts in building systematic reviews used in evidence-based medicine. We focus on clustering pre-selected classifiers based on a multi-measure prediction performance evaluation expressed in terms of a projection from a high-dimensional space to a visualizable two-dimensional one. The best classifier was selected from each cluster and included in the committee. We applied the committee of classifiers to rank biomedical abstracts based on the predicted relevance to the topic under review. We identified a subset of abstracts that represents the bottom of the ranked list (predicted as irrelevant). We used False Negatives (relevant articles mistakenly ranked at the bottom) as a final performance measure. Our early experiments demonstrate that the classifier committee built using our new approach outperformed committees of classifiers arbitrary created from the same list of pre-selected classifiers.

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

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.

Previous work by Kouznetsov et al. [2] proposed an algorithm to reduce the workload of building SRs while maintaining the required performance of the existing manual workflow.

Since the approach in [2] is based on using a committee of classifiers to rank biomedical abstracts based on the predicted relevance to the review topic, selecting

A. Farzindar and V. Keselj (Eds.): Canadian AI 2010, LNAI 6085, pp. 299–303, 2010.

the right classifiers to be included in the committee could be a key feature in improving prediction performance.

In this work we propose an approach to selecting classifiers based on a Projection-Based Framework for Classifier Performance Evaluation with respect to Multiple Metrics and Multiple Domains [3], [4], [5].

2 Method

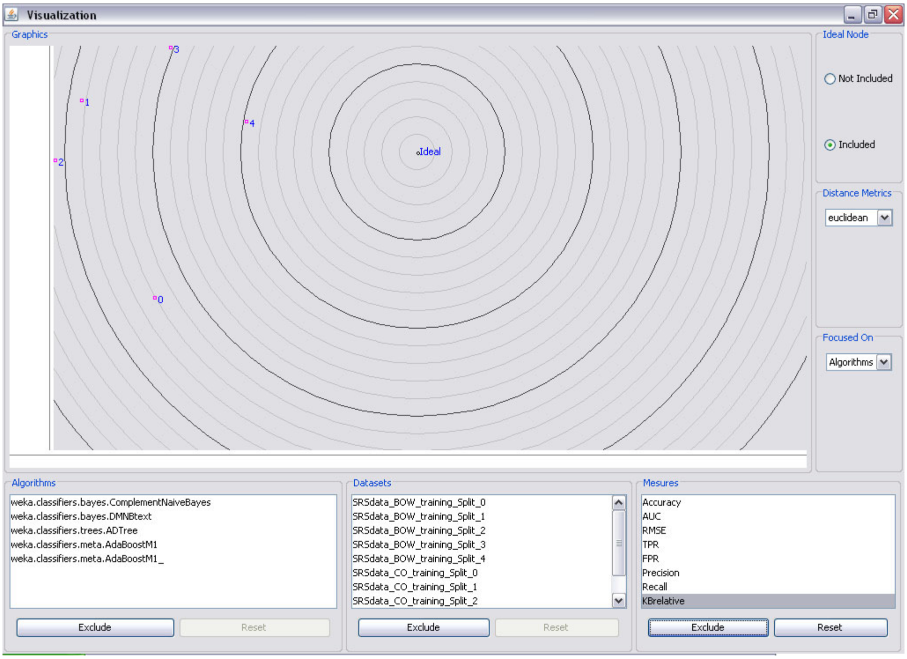
Ranking Method. We used the Ranking Algorithm, presented in the work of Kouznetsov et al. [2]. This approach is based on using committees of classification algorithms to rank instances based on their relevance to the topic of the review. It is a two-step ranking algorithm. While the first step, called 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. Finally, we get the collective rank which is assigned to each article in the test set. An instance with a higher collective rank is more relevant to the topic under review than another instance with a lower collective rank.

The classification decision of the committee is based on the collective ordered list of instances. The work in [2] provides ML techniques to establish the bottom threshold (number of instances at the bottom to be classified as negative with respect to the required level of prediction confidence). The articles with rank below the bottom threshold are predicted as not relevant and excluded from the review (in an attempt to reduce workload).

Visualizing method. Applying a ranking method assumes having a committee of classifiers that first needs to be selected. We propose a method for doing so that uses a **Visualizing Classifier Performance Tool (VCPT)** that was previously introduced in [3], [4], [5] and integrated with WEKA [7] in the context of this study. VCPT [9] includes our modification of the WEKA package to extend the functionality of its Experimenter Module integrated with the R Statistical package.

VCTL implements the following pipeline: All the classifiers involved in the study are run on all the data sets considered. The performance measures associated with one classifier on each dataset are organized into a single vector (any involved data set as well as any involved measure would produce a new dimension). The *Multidimensional Scaling* MDS [8] projection method (with Euclidean distance measure) is used to project high dimensional vectors into two dimensional vectors. Finally, projected vectors are visually presented on the VCTL graphical panel and users can visually separate classifiers into a few different clusters based on the performance achieved by each algorithm. We consider as a possibly good combination one where the best classifier is selected from each cluster and included in the team.

The visualization method we applied is a special case of the method proposed in the work of Japkowicz et al [5]. The purpose of this approach is to summarize the results obtained by classifiers on different data sets, using various performance measures. However, rather than summarizing these results numerically, we do so visually, exploiting the human visual skill, in the hope of retaining more information than we would by using a numerical summary.



Ideal Node
○ Not Included
● Included

Distance Metrics
euclidean

Focused On
Algorithms

Algorithms
weka.classifiers.bayes.ComplementNaiveBayes
weka.classifiers.bayes.DMNBtext
weka.classifiers.trees.ADTree
weka.classifiers.meta.AdaBoostM1
weka.classifiers.meta.AdaBoostM1_

Datasets
SRSdata_BOW_training_Split_0
SRSdata_BOW_training_Split_1
SRSdata_BOW_training_Split_2
SRSdata_BOW_training_Split_3
SRSdata_BOW_training_Split_4
SRSdata_CO_training_Split_0
SRSdata_CO_training_Split_1
SRSdata_CO_training_Split_2

Mesures
Accuracy
AUC
RMSE
TPR
FPR
Precision
Recall
KBrelative

Fig. 1. Visualizing Classifiers Performance with VCPT

3 Experiments

The work presented here was done on the same data that was used in [2]. The source data includes 23,334 medical articles pre-selected for the review topic "Assessment of health care delivery and outcomes in children and youth with special needs". The data contained only article titles and abstracts. While 19,637 articles have a title and an abstract, 3,697 articles have only a title. The data is available at http://www.site.uottawa.ca/~stan/public/

A stratified repeated random sampling scheme was applied to validate the experimental results. The data was randomly split into a training set and a testing set five times. On each split, the training set included 7,000 articles (626 positive and 6,374 negative) or about 30% of the entire data set, while the testing set included 16,334 articles (1,461 positive and 14,873 negative) or about 70% of the entire data set.

We applied two data representation schemes to build document-term matrices: the Bag-of- words (BOW) and the Second Order Soft Co-Occurrences (SOSCO) [6] representations.

The five classifiers included in the Classifier Committee used in [2] are considered as our pre-selected short list[1], namely: (0) Complement Naïve Bayes (CNB); (1) Discriminative Multinomial Naïve Bayes (DMNB); (2) Alternating Decision Tree (ADT); (3) AdaBoost [Logistic Regression] (AB/LR); (4) AdaBoost [j48] (AB/J48)];

[1] We are using this list just to illustrate the concept of building a classification team over a pre-selected set. In a real world setting, we would expect a wider list of algorithms, such as 10-20 algorithms.

Visualizing Classifier Performance Experiments. We ran classifiers on both the BOW data representation and the SOSCO data representation. Therefore we have 10 data sets (5 BOW data sets+5 SOSCO data sets). Since we used the 10 fold cross validation scheme each 7000-article set was randomly split into a new training set (6,300 instances) and a new testing set (700 instances), following a stratified strategy. The following eight performance evaluation metrics were included in each experiment: Accuracy, AUC, RMSE, True Positive Rate, False Positive rate, Precision, Recall, Kononenko & Bratko Relative Information score.

Committee Validation Experiments. The objective of these experiments is to compare the performance of committees of classifiers selected with VCPT against the performance of other possible committees. In these experiments, we are using both the training sets (7,000 articles on each split) and the testing sets (16,334 articles on each split). We consider the bottom threshold as 4000 and run the ranking algorithm to predict 4000 not relevant (negative) articles on the testing set. The paired two tailed t-test was applied to validate the statistical significance of the obtained results.

4 Results

The VCPT visual panel output is presented in Figure 1. We visually observed the obtained output and divided the classifiers into 3 clusters. The clusters are presented in Table 1.

We selected the following classifiers: Complement Naïve Bayes, Discriminative Multinomial Naïve Bayes, Ada Boosted J48. We call this committee the 0-1-4 team.

In order to evaluate our approach, we built two validation teams that include the same number of members: (1) Validation team 2-3-4 includes: Alternating Tree, Ada Boosted Logistic Regression, Ada Boosted J48; (2) Validation team 1-2-3 includes: Discriminative Multinomial Naïve Bayes, Alternating Tree,Ada Boosted Logistic Regression.

We built the validation committees based on a different approach from the one that yielded committee 0-1-4. While committee 0-1-4 includes only one classifier from every cluster, both committees 2-3-4 and 1-2-3 include two classifiers taken from the same cluster. These validation committees also used the best classifiers in order to maximize their performance.

Table 2 demonstrates the performance obtained with the ranking method by each classification committee. The performance measure is the number of False Negatives (FN) (taken as averages over 5 splits) that occurred for each classification committee over 4000 bottom ranked articles. False Negatives mean articles that are relevant to the SR topic but mistakenly ranked at the bottom of the ranking list.

The results we obtained have demonstrated that the committee completed with the proposed approach (0-1-4 team) outperformed both validation committees on the negative tail, namely the average FN of Team 0-1-4's output is around 48% less than the average FN of Team 2-3-4 (FN 11.2 vs. FN 21.4) and the average FN of Team 0-1-4's output is around 45% less than the average FN of Team 1-2-3 (FN 11.2 vs. FN 20.2). The applied t-test demonstrates that the achieved differences are statistically significant.

Table 1. Clusters visually built from VCPT outputs

Cluster #1	Cluster #2	Cluster #3
(3) AB/LR, (4) AB/J48	(1) DMNB, (2) ADT	(0) CNB

Table 2. False Negatives on Negative prediction zone (means over 5 folds)

Team 0-1-4	Team 2-3-4	Team 1-2-3
11.2	21.4	20.2

5 Conclusion and Future work

Our results demonstrate that using our approach can improve on the performance of classifier committees employed on Systematic Reviews. Based on our early experiments' output, the classifier committee formed by applying the projection method of classifier evaluation significantly overperformed the validation committees that consist of the same number of algorithms arbitrary included from the same list of pre-selected classifiers.

As a possible topic for future work we are planning more experimentation on different SR data sets and larger numbers of classifiers, to test whether our method can scale up.

References

1. Sackett, D., Rosenberg, W., Gray, J., Haynes, R., Richardson, W.: Evidence based medicine: what it is and what it isn't. BMJ 312 (7023): 71-2. PMID 8555924 (1996)
2. Kouznetsov, A., Matwin, S., Inkpen, D., Razavi, A., Frunza, O., Sehatkar, M., Seaward, L., O'Blenis, P.: Classifying Biomedical Abstracts Using Committees of Classifiers and Collective Ranking Techniques. In: Canadian Artificial Intelligence Conference (2009)
3. Alaiz-Rodriguez, R., Japkowicz, N., Tischer, P.: Visualizing Classifier Performance. In: Proceedings of the 20th IEEE International Conference on Tools for Artificial Intelligence, ICTAI 2008 (2008)
4. Alaiz-Rodriguez, R., Japkowicz, N., Tischer, P.: A Visualization-Based Exploratory Tool for Classifier Comparison with respect to Multiple Metrics and Multiple Domains. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part II. LNCS (LNAI), vol. 5212, pp. 660–665. Springer, Heidelberg (2008)
5. Japkowicz, N., Sanghi, P., Tischer, P.: A Projection-Based Framework for Classifier Performance Evaluation. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part I. LNCS (LNAI), vol. 5211, pp. 548–563. Springer, Heidelberg (2008)
6. Razavi, A.H., Matwin, S., Inkpen, D., Kouznetsov, A.: Parameterized Contrast in Second Order Soft Co-Occurrences: A Novel Text Representation Technique in Text Mining and Knowledge Extraction. In: Second International Workshop on Semantic Aspects in Data Mining (SADM 2009), USA, Miami (2009)
7. Software package Weka, http://www.cs.waikato.ac.nz/ml/weka/
8. Cox, T., Cox, M.: Multidimensional Scaling. Chapman and Hall, Boca Raton (October 1994)
9. Visualization Software for Clasifier Evaluation, http://www.site.uottawa.ca/~nat/Visualization_Software/visualization.html