Building comprehensive searches through a Machine Learning approach for systematic reviews

Corrado Lanera[1] (corrado.lanera@unipd.it), Clara Minto[1] (clara.minto@unipd.it), Abhinav Sharma[2] (abhishar@iitk.ac.in), Dario Gregori[1] (dario.gregori@unipd.it), Paola Berchialla[3] (paola.berchialla@unito.it), Ileana Baldi[1] (ileana.baldi@unipd.it).

[1] Unit of Biostatistics Epidemiology and Public Health (UBESPH) - Department of Cardiac, Thoracic and Vascular Sciences (DCTV) - University of Padova (Italy. [2] Biological Sciences and Bioengineering (BSBE) Department - IIT Kanpur (India). [3] Department of Clinical and Biological Science (DSCTB) - University of Torino (Italy)

Background and aims

Clinical trial registries are important literary sources that contribute to a medicine based on up-to-date evidence.[1] Despite their significant role, their use is not frequent in systematic reviews and meta-analyses, probably because of difficulties in managing records.[2]

Main purpose: Our goal was to propose an automated approach to identify relevant information on a clinical trials registry. We used text mining and machine learning techniques to train a classifier able to detect relevant records on registry database, once a search has already been performed on standard data source (e.g. PubMed)

Challenges: Ensure that text data can be used to train and predict on different data sources taken in different time and from different kind of databases. Also, be able to deal with highly unbalanced sets of data, with which most classifiers are biased towards the majority class performing poorly on minority ones.[3]

Motivating example(s): The 14 Systematic reviews and Meta-analyses re-analyzed by BMJ. [1]

Methods: We explored the combinations of five balancing methods (including unbalanced and randomized sampling with both 50:50 and 35:65 Negative Records) and four MLTs (within GLMNet, kNN, SVM, and RF). One Systematic Review was used to select the best combination to implement in general. The best one was later used for a pool of fourteen studies for which both original and improved (by BMJ) results are available. For each study, the selected technique was trained on Positive PubMed records and on a series of random negative records all "out of" the PubMed string search used by the Sistematic Reviews' authors, with a 1/20 ratio up to a multiple of 200 records. The text was processed by lowering, removing the numbers and punctuations and augmented with the bi-gram tokens. The trainings were 10 Fold Cross-Vlidated for each of value of a set of three random parameters for the MLT and each one included re-weighting processes with TF-iDF, feature selection (down to 4% of the top TF-iDF records) and balance. The best validated parameter was used to teach the final classifier, tested on an independent set consisting of the records from ClinicalTrial.gov proposed by Baudard et al. and a set of negative random records (different for each study) to form a 100 records test-set for each study. Validation and test DTMs have been adapted each time on the specifications requested by the applied machine.

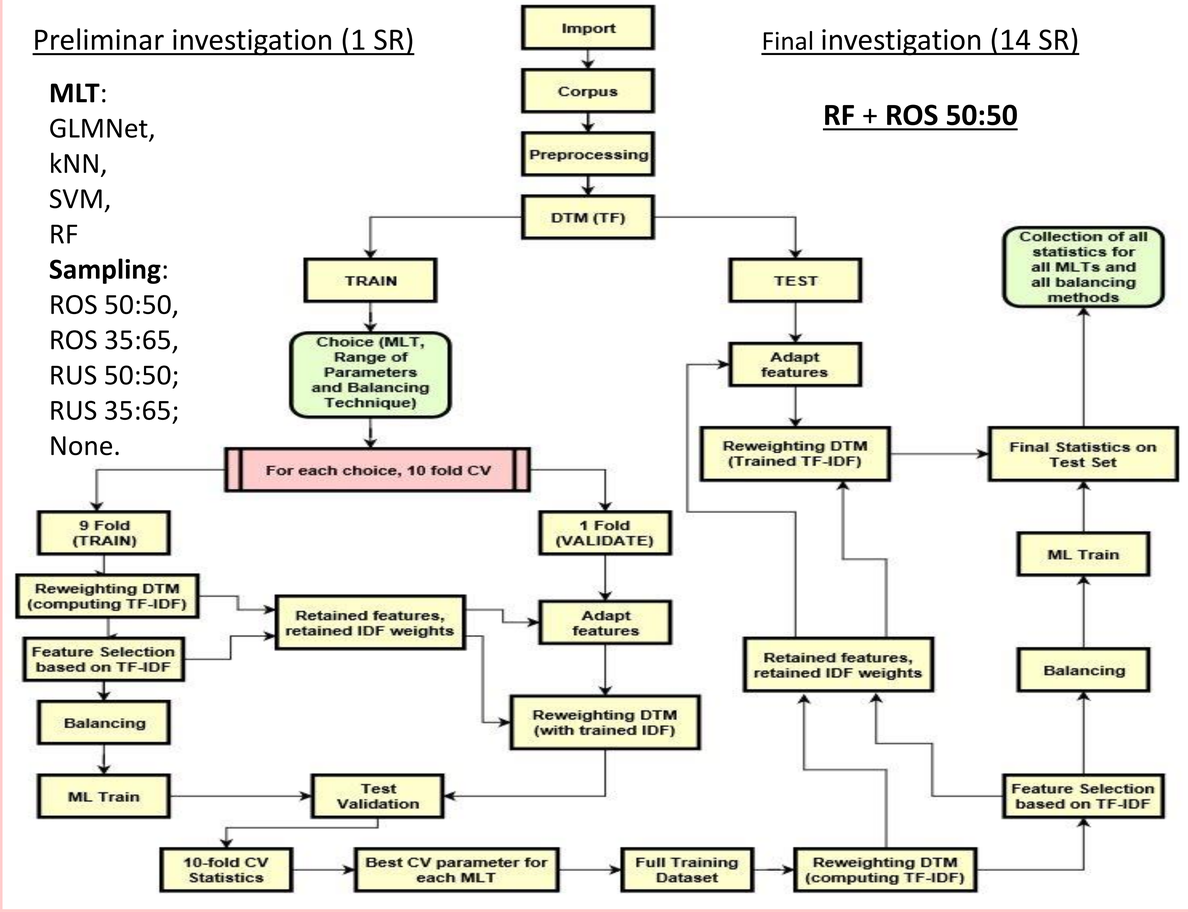
Computation Plan

Preliminar investigation (1 SR)

MLT:
GLMNet,
kNN,
SVM,
RF
Sampling:
ROS 50:50,
ROS 35:65,
RUS 50:50;
RUS 35:65;
None.

Final investigation (14 SR)

RF + ROS 50:50



Results

Median AUC	0.798
Median SENS & RECALL	1
Median SPEC	0.596

	sr	train_pos	train_neg	test_pos	test_neg	AUC	PPV	PREV	SENS	SPEC
Yang	4	200	1	99	NA	NA	NA	NA	NA
Meng	9	200	1	99	0.677	0.015	0.01	1	0.354
Segelov	13	400	2	98	0.770	0.043	0.02	1	0.541
Li	6	200	1	99	0.859	0.034	0.01	1	0.717
Lv	12	400	1	99	0.798	0.024	0.01	1	0.596
Wang	32	800	1	99	0.909	0.053	0.01	1	0.818
Zhou	9	200	1	99	0.803	0.025	0.01	1	0.606
Liu	23	600	21	79	0.918	0.618	0.21	1	0.835
Douxfils	13	400	1	99	0.606	0.013	0.01	1	0.212
Kourbeti	75	1600	4	96	0.854	0.125	0.04	1	0.708
Li	9	200	2	98	0.592	0.024	0.02	1	0.184
Cavender	6	200	1	99	NA	NA	NA	NA	NA
Chatterjee	18	400	1	99	0.515	0.010	0.01	1	0.030
Funakoshi	9	200	2	98	NA	NA	NA	NA	NA

Discussion

The proposed procedure has led to the identification, for each systematic review, of all the additional records identified by the Baudard et al. study, and also leads to discreetly high AUC values as a proof of non-triviality. Positive Predictive Values (PPV) have been reported despite they are highly biased by the arbitrarily small prevalence of positive cases into the test set.

The validity of application of Text-Mining techniques in synergy with Machine Learning ones to meet the needs of researchers involved in systematic reviews had already been positively examined.[4] Our study solves, first, the limitation highlighted above of a limited number of different test studies tested, on the other hand, not only improves the previous performance but also exerts the applicability not only of reducing manual labor but also of using very important but scarcely used resources so far.

References

[1]Baudard, Marie, et al. "Impact of searching clinical trial registries in systematic reviews of pharmaceutical treatments: methodological systematic review and reanalysis of meta-analyses." *bmj* 356 (2017): j448.

[2]Jones CW, Keil LG, Weaver MA, Platts-Mills TF. Clinical trials registries are under-utilized in the conduct of systematic reviews: a cross-sectional analysis. *Systematic Reviews*. 2014;3:126. doi:10.1186/2046-4053-3-126.

[3]Longadge R, Dongre S. Class imbalance problem in data mining review. ArXiv Prepr ArXiv13051707 [Internet]. 2013; Available from: https://arxiv.org/abs/1305.1707

[4]Bekhuis T, Demner-Fushman D. Screening nonrandomized studies for medical systematic reviews: a comparative study of classifiers. *Artificial Intelligence in Medicine*. 2012;55(3):197-207. doi:10.1016/j.artmed.2012.05.002.