Supervised Methods to Support Online Scientific Data Triage

Hayda Almeida[1], Marc Queudot[1], Leila Kosseim[2], and Marie-Jean Meurs[1]

[1] Université du Québec à Montréal
[2] Concordia University
Montreal, QC, Canada
meurs.marie-jean@uqam.ca

Abstract. This paper presents machine learning approaches based on supervised methods applied to triage of health and biomedical data. We discuss the applications of such approaches in three different tasks, and evaluate the usage of triage pipelines, as well as data sampling and feature selection methods to improve performance on each task. The scientific data triage systems are based on a generic and light pipeline, and yet flexible enough to perform triage on distinct data. The presented approaches were developed to be integrated as a part of web-based systems, providing real time feedback to health and biomedical professionals. All systems are publicly available as open-source.

1 Introduction

Scientific data processing is a very important and often critical step in the routine of life science practitioners. In certain cases, the ability to find crucial information quickly can be decisive to help patients, or to avoid bottlenecks in the scientific research workflow. Health care and biomedical professionals can make use of many online data sources, which can provide them with essential information. For example, scientific researchers constantly rely on online literature databases to support their work, while health care professionals have the opportunity of following up closely on patient conditions through their interactions on health forum posts. Going through these large sets of data to fetch critical pieces of information can be overwhelming. Systems that perform scientific data triage automatically can therefore be a powerful tool for health and biomedical professionals, since they can help with identifying relevant information faster in large datasets. Additionally, the automatic triage when integrated in web-based systems, can provide a data relevance feedback in real time for life-science professionals.

In this paper, we present three automatic approaches based on supervised machine learning to perform scientific data triage. Supervised methods make use of relevant data that was previously labeled by the subject experts to derive a pattern, used as a model. This model is applied to predict the relevance of new data that needs to be analyzed. We present three supervised learning approaches for automatic triage of biomedical and health data, that are suitable for web-based systems. First, we describe an approach to select scientific literature related to fungal enzymes that can be used in bioproduct conversion processes. Second, an approach to support scientific literature screening for HIV systematic reviews is described. Finally, we present an on-going project to predict the severity of user posts in the ReachOut mental health forum. These approaches were developed using open access data, and are based on light and generic methods. Their design allow for their usage on various data, as well as their integration in online platforms.

2 Related Work

Systems developed to support health and biomedical practitioners are important tools to help dealing with large amounts of scientific data in a responsive and efficient manner [15, 8]. Health and biomedical data available online represent a highly valuable, often essential, resource in the routine of science professionals and researchers [14, 9].

Data triage systems can assist these professionals in acquiring knowledge from scientific data, allowing them to save time, scale up their work, and reduce bottlenecks in their knowledge discovery or response workflow [20, 16]. More importantly, health and biomedical triage tools can be decisive to keep up with the quick pace of scientific research, and the need of timely responses in health care when data are found in online platforms. Studies have investigated several approaches to better performing scientific data triage. Machine learning methods have been applied to handle large biological datasets [21], sort scientific literature in systematic review workflows [5, 21], help identify mental health issues [18], and recognize user sentiments [19]. Previous works have evaluated the importance of integrating intelligent systems to web-based environments to assist scientific data processing. Their applications are numerous, such as heart disease automatic classification [17], literature mining [12], or even finding patterns in bioimages [10]. In this context, specific strategies to process data can be substantially beneficial for integrating supervised learning methods into online platforms. Methods such as data sampling [7] and feature selection [11] have been shown to reduce the computational cost of learning tasks, while still maintaining or even improving performance [22, 4].

Here, we describe the methods implemented to tackle the scientific data triage in three different tasks, using supervised machine learning, data sampling and feature selection. The approaches were developed to be integrated in web-based systems.

3 Proposed Triage Approaches

We describe here three approaches proposed to handle the task of health and biomedical data triage. The tasks presented are (1) *Discovery of fungal enzymes*, (2) *HIV systematic review*, and (3) *Response urgency of mental health forum posts*. These approaches were developed to meet the specific requirements of each application, such as utilizing the most fitting feature set, and handling different data inputs. However, the presented systems are based on a generic pipeline and architecture. This allows them to be used in other applications and domains. Such approaches can not only be applied in tasks handling data regarding new topics, but can also be easily integrated in online and real-time systems.

The pipeline of the data triage systems is based on a supervised machine learning approach, which is divided in two phases: training and testing. Before being able to provide relevance predictions for new data, the triage systems have to be trained based on manually labeled corpora. Documents in the training corpora were labeled according to their relevance for health and biomedical practitioners, given a certain topic. In task (1), documents were labeled as either *positive* or *negative*. In task (2), documents were labeled as either *included* or *excluded*. For these cases, *positive* or *included* documents were the ones of potential interest. In task (3), documents were labeled as *crisis*, *red*, *amber* or *green*, to indicate the urgency of a forum moderator to intervene. After being trained, the systems are capable of outputting predictions for new data. The data triage systems presented here were applied in different tasks. However their architecture follows the same pipeline, divided into four lightweight and generic main modules:
Data Handling: this module is responsible for processing the data, by handling all input documents, labeled or unlabeled. During this step, selected content from documents is gathered, after going through normalization processes.
Feature Extraction: at this step, discriminative features are extracted from the training data. Multiple feature extraction methods can be applied, and they can be used in an combined manner, creating a feature set. The usage of different features yields distinct performances depending on the task subject. Additionally, features can also go through a filtering process, where feature selection algorithms are used to identify which features seem more discriminative for a given task.
Model Building: this module represents the task data in terms of the feature set chosen. At this step, a model is generated after processing the features representing the data, using a classification algorithm. The model generated by the training data and selected feature set is what provides knowledge to the triage systems, since it depicts the underlying pattern in the task data.
Document Predictor: the prediction module relies on the model generated by the *Model Building* module. Once new unlabeled input data are represented in the same manner as the training data (using the same feature set), the triage systems make use of the generated model to output relevance predictions regarding these new data.

The data triage approaches can be easily integrated in online platforms, and provide real-time feedback for new input data. Each triage system was incorporated in web-based applications, to support professionals in going through large volumes of health and biomedical data. In the following Sections, we provide further details regarding the approach adopted in each of the three triage tasks.

3.1 Discovery of Fungal Enzymes

The task of discovering fungal enzymes [2] aimed at selecting relevant literature to support the identification of fungal enzymes used in industrial processes. To identify the most fitting set-up for this task, over 100 classification models were designed and evaluated, using 4 feature settings, 3 supervised learning algorithms, and 9 differently balanced corpora.

The feature settings were composed of Enzyme Commission (EC) numbers, Bag-of-Words (BOW) representation of the document content, and annotated bio-entities, which were extracted from documents with the help of the mycoMINE [13] annotation tool. The different corpora were generated based on a data sampling technique, applied to study the usage of corpora with more equally balanced class distributions. The three algorithms used in this task were Naïve Bayes (NB), Logistic Model Trees (LMT), and Support Vector Machine (SVM).

The system for discovery of fungal enzymes was created to be integrated with the Proxiris [6] web-based tool, which supports scientific data mining by identifying entities of interest in web literature. Combining both approaches facilitates the discovery process of candidate fungal enzymes, since users can visualize entities of interest in the literature, and have a relevance prediction about the whole document in real-time.

3.2 HIV Systematic Review

The process of systematic reviews entails finding studies that can answer a given research problem. The HIV systematic review triage [3] was developed to support researchers working on the SHARE [1] database in identifying potentially relevant literature.

Over 100 classification models were analyzed to identify the most appropriate configuration to address the problem, using 3 feature types, 2 feature selection methods, 5 differently balanced corpora, and 3 classification algorithms: NB, LMT, and SVM.

The different corpora used in this task were also generated based on sampling techniques, to analyze the results of training based on various class distributions. The feature sets were composed of different combinations of a BOW representation of the document content, SHARE experts selected keywords, and Medical Subject Headings[2] (MeSH) terms found in documents. Two methods were used to filter out features according to their discriminative power~~,~~ : Inverse Document Frequency (IDF) and Odds Ratio (OR).

The HIV systematic reviews triage was designed to be integrated in the SHARE web-based tool, to help researchers in the process of quickly determining if new documents should be reviewed by SHARE curators.

[1] http://www.hivevidence.ca

[2] https://www.ncbi.nlm.nih.gov/mesh

3.3 Response Urgency of Mental Health Forum Posts

The third system, for triage of mental health forum posts [1], was developed for the CLPsych Shared Task[3]. The goal of the system is to guide forum moderators to quickly assess the response urgency required from a post, given the content that the user has posted.

A specific normalization method was created to handle the forum data. User posts contained images, URLs, and emoticons, which were replaced by a corresponding word, since they could be helpful to indicate the sentiment related to a post.

The feature set used in this task was made of bigrams (consecutive word pairs), Part-of-Speech (POS) tags, and sentiments. The vocabulary of two sentiment libraries was selected to annotate post sentiments, while the POSTaggerAnnotator [4] was used to annotate POS tags in posts. All feature types were filtered using a Correlation-based Feature Selection (CFS).

The automatic classification of posts was performed by Bayesian Network (BN), Sequential Minimal Optimization (SMO), and LMT algorithms. A rule-based classification method was merged with the automatic classification output to finally provide risk predictions for each post. The rules were based on a discriminative vocabulary selected from the least represented (and more urgent) post labels: *red* and *crisis*.

One of the CLPsych task interests is to integrate the proposed techniques in the web forum ReachOut[5], so forum moderators can assess in real time if user posts need to be treated urgently.

4 Experimental Results

We present here the summary of results for each of the three tasks described in this paper. Since over 100 models were evaluated for some of these tasks, we will only present the models that yielded the best performances, in addition to an explanation of the best result obtained on each triage.

Discovery of Fungal Enzymes The experiments ran for the discovery of fungal enzymes (see Section 3.1) evaluated the usage of different balance ratios for the training corpora (using various ratios of *positive* labeled documents), classification algorithms, and feature combinations.

The performance is shown for the least represented *positive* label. Table 1 shows that the best results were achieved with training corpora presenting a balanced distribution of labels, the LMT and SVM classification algorithms, and either the BOW representation of documents or all the features combined. The results are ranked by F-2 score, a generalization of the F-measure (harmonic mean of Precision and Recall) using twice the weight for Recall than Precision.

[3] http://clpsych.org/shared-task-2016/

[4] http://nlp.stanford.edu/software/tagger.shtml

[5] http://au.reachout.com/

The model that yields the best performance is composed of the LMT classifier, using a balanced dataset (containing 40% of *positive* documents), and represented by a combination of all the feature types. Another model that yields similar performance, is composed of the SVM classifier, a balanced dataset (containing 35% of *positive* documents), and BOW features.

Positive label %	Feature set	Classifier	Precision	Recall	F-m	F-2
40%	all features	LMT	0.361	0.847	0.506	0.670
30%	all features	LMT	0.398	0.780	0.527	0.650
35%	BOW	SVM	0.369	0.800	0.505	0.650
35%	BOW	LMT	0.359	0.807	0.497	0.650
35%	all features	SVM	0.357	0.793	0.493	0.640

Table 1. Task (1): Best Results on the Discovery of Fungal Enzymes

HIV Systematic Review The triage for HIV systematic reviews (see Section 3.2) also evaluated training data sampling, along with a comparison between the use of SHARE keywords, BOW or BOW with MeSH terms as features. The usage of BOW and MeSH terms was also analyzed before and after IDF and OR feature selection methods. Table 2 reports the results obtained with LMT and SVM models, which demonstrated better performance.

Models using OR as feature selection used ≈80% less features than models without any feature selection, and yet yielded comparable results. Also in this task, performance metrics are shown for the least represented *included* label.

The best model is composed of the LMT classifier, a balanced dataset (containing 40% of *positive* documents), and no feature selection. The model composed of the same configurations, but utilizing OR as feature selection, achieves similar performance, but the process considerably less data after filtering out features.

Positive label %	Feature set	Feature selection	Classifier	Precision	Recall	F-m	F-2
40%	BOW + MeSH	N/A	LMT	0.467	0.900	0.615	0.759
40%	BOW + MeSH	OR	LMT	0.445	0.882	0.591	0.737
30%	BOW	N/A	SVM	0.540	0.800	0.645	0.730
30%	BOW	OR	SVM	0.497	0.827	0.621	0.730

Table 2. Task (2): Best Results on HIV Systematic Review

Response Urgency of Mental Health Forum Posts The approaches that best suited the triage of mental health forum posts were chosen after performing experiments evaluating different techniques. At first, a relevant set of features was selected, based on the CFS method. The relevant features were used to generate a supervised classification model, and get predictions on the data. Additional predictions were obtained through a rule-based classification approach. Finally, the supervised and rule-based classification predictions were merged to produce the final predictions.

Performance is presented in terms of the official CLPsych metrics, which included accuracy and macro-averaged F-score (macro F-m). The task metrics are based on the system capability of highlighting the labels of higher interest (*crisis*, *red*, and *amber*). Table 3 shows that the best results were achieved using the merged approach, using SMO and a set of 5 rules, and LMT using a set of 3 rules.

Approach	macro F-m	accuracy	non-green v. green macro F-m	non-green v. green accuracy
SMO + 5 rules	0.29	0.74	0.68	0.82
LMT + 3 rules	0.27	0.72	0.72	0.83
LMT + 5 rules	0.26	0.72	0.72	0.83
LMT only	0.25	0.75	0.75	0.85

Table 3. Task (3): Best Results on Response Urgency of Mental Health Forum Posts

5 Conclusion

In this paper we presented an overview of three different approaches to perform triage of scientific data. We evaluated the performance of several supervised learning models by using a common data triage pipeline between all tasks. At the same time we applied specific methods to meet the requirements of each task, such as data sampling and feature selection. The scientific data triage pipeline is based on light and generic modules, allowing the systems to be used to process other data types, and to be integrated in online platforms.

Generally, models using the LMT classifier outperformed all other models. Models based on SVM, however, yielded similar performance to LMT models, and can be suitable if an even shorter response time is required. The usage of dataset sampling yielded better performance on tasks (1) and (2), and feature selection methods improved performance on tasks (2) and (3). Such techniques allow the classification models to perform well using less computational resources, providing lightweight solutions that can respond efficiently in real-time and online systems.

Ongoing work Novel classification approaches are being developed to improve the results on task (3). Along with evaluating the system with a new dataset, two different methods are currently under development and testing: ensemble classification and deep neural networks.
The systems presented in this paper are publicly available, and can be found at:
https://github.com/TsangLab/triage
https://github.com/TsangLab/mycoSORT
https://github.com/BigMiners/CLPsych2016_Shared_Task

References

1. H. Almeida and M.-J. Meurs. Automatic Triage of Mental Health Online Forum Posts - NAACL-CLPsych 2016 System Description. *Red*, 110(11.61):27, 2016.
2. H. Almeida, M.-J. Meurs, L. Kosseim, G. Butler, and A. Tsang. Machine Learning for Biomedical Literature Triage. *PLOS ONE*, 9(12):e115892, 2014.
3. H. Almeida, M.-J. Meurs, L. Kosseim, and A. Tsang. Data Sampling and Supervised Learning for HIV Literature Screening. *IEEE Transactions on NanoBioscience*, 15(4):354–361, June 2016.
4. T. Basu and C. Murthy. Effective Text Classification by a Supervised Feature Selection Approach. In *Proceedings of the IEEE 12th International Conference on Data Mining Workshops (ICDMW), December 10, Brussels, Belgium*, pages 918–925. IEEE, 2012.
5. T. Bekhuis and D. Demner-Fushman. Screening Nonrandomized Studies for Medical Systematic Reviews: A Comparative Study of Classifiers. *Artificial intelligence in medicine*, 55(3):197–207, 2012.
6. V. Chahinian, M.-J. Meurs, D. H. Mason, E. McDonnell, I. Morgenstern, G. Butler, and A. Tsang. Proxiris, an Augmented Browsing Tool for Literature Curation. In *Proceedings of 9th International Conference on Data Integration in the Life Sciences, Springer*, 2013.
7. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling TEchnique. *Journal of Artificial Intelligence Research*, 16:341–378, 2002.
8. A. Holzinger and I. Jurisica. Knowledge Discovery and Data Mining in Biomedical Informatics: The Future is in Integrative, Interactive Machine Learning Solutions. In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*, pages 1–18. Springer, 2014.
9. D. Howe, M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, D. P. Hill, R. Kania, M. Schaeffer, S. St Pierre, S. Twigger, O. White, and S. Yon Rhee. Big data: The Future of Biocuration. *Nature*, 455(7209):47–50, 2008.
10. J. Kölling, D. Langenkämper, S. Abouna, M. Khan, and T. W. Nattkemper. WHIDE - A Web Tool for Visual Data Mining Colocation Patterns in Multivariate Bioimages. *Bioinformatics*, 28(8):1143–1150, 2012.
11. H. Liu, H. Motoda, R. Setiono, and Z. Zhao. Feature Selection: An Ever Evolving Frontier in Data Mining. In *Proceedings of the 4th Workshop on Feature Selection in Data Mining, June 21, Hyderabad, India*, pages 4–13, 2010.
12. Z. Lu. PubMed and Beyond: A Survey of Web Tools for Searching Biomedical Literature. *Database*, 2011:baq036, 2011.
13. M.-J. Meurs, C. Murphy, I. Morgenstern, G. Butler, J. Powlowski, A. Tsang, and R. Witte. Semantic Text Mining Support for Lignocellulose Research. *BMC Medical Informatics and Decision Making*, 12(1):S5, 2012.
14. S. A. Moorhead, D. E. Hazlett, L. Harrison, J. K. Carroll, A. Irwin, and C. Hoving. A New Dimension of Health Care: Systematic Review of the Uses, Benefits, and Limitations of Social Media for Health Communication. *Journal of Medical Internet Research*, 15(4):e85, 2013.
15. T. B. Murdoch and A. S. Detsky. The Inevitable Application of Big Data to Health Care. *JAMA, The Journal of the American Medical Association*, 309(13):1351–1352, 2013.
16. A. O'Mara-Eves, J. Thomas, J. McNaught, M. Miwa, and S. Ananiadou. Using Text Mining for Study Identification in Systematic Reviews: A Systematic Review of Current Approaches. *Systematic Reviews*, 4(1):5, 2015.

17. S. Palaniappan and R. Awang. Intelligent Heart Disease Prediction System Using Data Mining Techniques. In *IEEE/ACS International Conference on Computer Systems and Applications, 2008.*, pages 108–115. IEEE, 2008.
18. S. Saleem, R. Prasad, S. N. P. Vitaladevuni, M. Pacula, M. Crystal, B. Marx, D. Sloan, J. Vasterling, and T. Speroff. Automatic Detection of Psychological Distress Indicators and Severity Assessment from Online Forum Posts. In *COLING, The International Conference on Computational Linguistics*, pages 2375–2388, 2012.
19. M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment Strength Detection for the Social Web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, 2012.
20. S. Tuarob, C. S. Tucker, M. Salathe, and N. Ram. An Ensemble Heterogeneous Classification Methodology for Discovering Health-related Knowledge in Social Media Messages. *Journal of Biomedical Informatics*, 49:255–268, 2014.
21. M. Wang, W. Zhang, W. Ding, D. Dai, H. Zhang, H. Xie, L. Chen, Y. Guo, and J. Xie. Parallel Clustering Algorithm for Large-Scale Biological Data Sets. *PLOS ONE*, 9(4):e91315, 2014.
22. G. M. Weiss, K. McCarthy, and B. Zabar. Cost-sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs? In *DMIN-International Conference on Data Mining*, pages 35–41, 2007.