Biomedical Literature Triage using Supervised Learning

Hayda Almeida[1], **Marie-Jean Meurs**[2]
[1]Department of Computer Science and Software Engineering
[2]Centre of Structural and Functional Genomics
Concordia University, Montréal, QC
h_marci@encs.concordia.ca
marie-jean.meurs@concordia.ca

Abstract

We present the ongoing development of a supervised machine learning approach to support the automatic triage of HIV-related literature. The system will be capable to learn from correctly classified samples of candidate PubMed abstracts, and then it will be able to provide a classification decision for a new document according to specific attributes found in the data. In this project, we analyze the discriminative power of different data features by experimenting with several feature extraction approaches. We also study the application of data sampling techniques, in an attempt to overcome the imbalance found in the dataset class distribution. Our goal is to define the most fitting classification model to predict the document relevance to the HIV-related research.

1 Introduction

The amount of documents currently found in biological research data repositories is massive, and this number is rapidly growing. As of December 2014, PubMed [1], the largest knowledge source for biomedical and life science literature, comprises more than 24 million citations, including MEDLINE, life science journals and online books. The number of scientific publications related to HIV research is also very high.

If a user queries only the keyword *HIV* to retrieve related publications on PubMed, the list of results will contain almost 300,000 occurrences. When analyzing the volume of publications over time, it is possible to observe the sound increase of publications in this field: less than 10,000 HIV related articles were added to PubMed during the year 2000, while in 2014 had over 16,200 more HIV related articles.

Facing such a vast universe of available data, the task of identifying relevant articles for a specific research domain can be compared to looking for a needle in a haystack. The information filtering process performed by researches can turn into a time-consuming and error-prone activity. This scenario is a normal routine experienced by biocurators when trying to discover new and relevant research data.

Curators will usually collect a long list of potentially interesting documents when querying biomedical databases for keywords. Only after each document abstract is analyzed, the curators can decide about rejecting or moving the document forward in the manual curation workflow. This selection step, called literature triage, represents a bottleneck in the curation knowledge discovery workflow [2].

1.1 Project relevance

Research programs dedicated to study public health generally need to manipulate and analyze large datasets, as well as perform systematic reviews of biomedical literature [3]. The capability of performing automatic classification of paper abstracts can be meaningful for HIV-related or even other research programs, because it would directly affect the coverage and quality of knowledge discovery.

Substantial efforts are put into extracting and annotating information on life science related documents [4, 5], with the use of natural language processing approaches [6]. Several tools have been developed to help biology researchers and literature curators to identify relevant data in the extensive available literature [7, 8, 9, 10, 11]. Most of these works aimed to improve the visualization of relevant terms or key-words in documents. This could help curators to more quickly identify the potentially relevant among all papers returned from a search.

However, these term-based tools are not yet capable of supporting curators to determine the relevance of a certain document according to the research triage criteria. Relevance-based approaches suited to perform this filtering by automatically sorting document abstracts are still being developed [12].

The task of biomedical literature triage has a fundamental characteristic of imbalance in the classes distribution. Since curators usually reject the majority of documents retrieved after a database search, the number of negative (non-curatable) documents is much higher than the amount of positive (curatable) documents encountered.

In this project, we will investigate strategies to overcome the imbalanced class distribution on a biomedical dataset, as well as study discriminative features to best develop a supervised learning model capable of supporting the triage of HIV-related documents.

2 Related work

Designing a supervised learning model that is capable of supporting the triage of biomedical literature can be challenging. A dataset that can fairly represent the realistic class distribution of this task will present a strong imbalance between relevant (curatable) and non-relevant (rejected) class labels among the document instances.

Datasets with imbalanced class distributions are commonly found in a variety of fields besides biomedical text classification. Some areas of study that usually deal with this specific issue are as speech recognition [13], medical diagnosis [14] [15], and fraud and image detection [16][17].

The class imbalance greatly affects the classifier performance, as explained by [18]. This happens because non-relevant instances are massively represented in the dataset if compared to the amount of instances belonging the relevant class. Therefore, the classifier model has many more examples of the majority class to learn information from, and this can introduce a bias in the prediction process.

The imbalance dataset issue has been studied and pointed out as an important condition for supervised learning tasks [19] [20] [21] [22] and various approaches have been evaluated to overcome it.

Cost-sensitive classifiers and data-sampling are two common methods that were studied as an alternative to tasks that present class imbalance. The former method is implemented at the algorithm level, while the latter is implemented at the data level. The strategy used by cost-sensitive classifiers [23] is to lower classification errors that occur in the minority class by intentionally introducing a bias, such a weight, so the classification errors made in the minority class are viewed as more costly than errors made in the majority class.

Data sampling methods are presented and discussed by [24]. The technique described by the author, Synthetic Minority Over-sampling Technique (SMOTE), covers two sampling strategies: under-sampling and over-sampling. Under-sampling consists in reducing the number of instances in the majority class. Over-sampling tries to increase the number of instances in the minority class by adding new synthetically generated instances.

A comparison study [25] attempted to compare cost-sensitive and data sampling when handling imbalance datasets. The classification results of both strategies could not clearly indicate that one method outperformed the other.

When evaluating the sampling strategy, [23] and [26] pointed that its performance is comparable to other state-of-the-art strategies. However, when compared to cost-sensitive, the sampling strategy is understood as a less restrictive method [25]. The fact that sampling is limited to the data level of the learning process, makes it a more flexible strategy to be used across different types of tasks, if compared to the cost-sensitive approach. A cost-sensitive classifier, that has to implement changes at the algorithm level, could be restrictive in certain types of tasks.

Regarding the imbalanced class distribution, the work of [25] recommended under-sampling to also handle classification problems in which the dataset is larger than the computational power to process it or to reduce the time spend in the training phase.

3 Methodology

Corpus and Data Sampling. The dataset used to train the system will be composed of the same ratio of relevant/non-relevant document instances as the one faced by the curators when performing manual triage. We will experiment with a sampling technique [24], by under-representing the majority class in order to handle the imbalanced dataset class distribution.

Feature extraction Relevant fragments of text will be extracted from the documents to be used as features in our classification models. Features will be mostly selected from the "AbstractText" and "ArticleTitle" text fields. The MeSH terms [27] annotated in the document will be extracted from the "MeshHeadingList" text field.

A feature vector will be built to represent each document in the collection as the occurrence of features in its text, and its classification label. All vectors will be combined in a matrix that represents the complete dataset of the task. The large size of the dataset leads to a sparse and also large representation matrix, which demands high computational processing capabilities.

We will study different techniques to reduce the feature space size, by experimenting with feature selection methods. As per a first iteration, we will evaluate the classification performance after filtering by feature character length and feature occurrence. By applying such filters, words found only once in the entire training corpus, or words that contain no more than 3 characters will not be considered to generate feature vectors.

Classification Algorithms. Three different classification algorithms will be used in our experiments: Naïve Bayes (NB), Logistic Model Trees (LMT) and Support Vector Machine (SVM).

The experiments with the NB classifier will be used as a baseline evaluation of the sampling and feature selection strategies applied in the different models.

The NB classifier is a probabilistic model based on Bayes' Rule. It assumes a strong conditional independence of features, building a "naïve" independence model. NB considers that in a feature vector F, the features $F_1, ..., F_n$ are conditionally independent from each other, given a class C. By this assumption, Naïve Bayes implies that the presence of one word (feature) is not correlated with the presence or absence of another word in a document, given a class label.

The LMT algorithm [28] was reported as an efficient classifier to handle tasks with imbalanced datasets [29]. LMT is a combination of Decision Tree and LogitBoost algorithms. It is formed by a classification tree, with logistic regression models on its nodes. In the decision tree nodes, the LogitBoost algorithm trains a subset of data for a number of iterations to define a logistic regression model for the current node. To split the current subset, a Decision Tree criterion is applied.

SVM [30] was previously indicated as an algorithm capable of dealing with imbalanced data [31, 32, 33]. The SVM model considers the "margin maximum classifier" [**?**], which consists in the largest radius around a classification boundary, and tries to separate data points on a dimensional space to identify the different classes.

Evaluation Metrics. The experimental results will be evaluated according to five different metrics: Precision, Recall, F-measure, F-2 and Matthews Correlation Coefficient. Precision computes the

number of correct predictions between all correct and incorrect predictions made by the classifier for a specific class. It evaluates the ability of a classifier to generate relevant outputs.

Recall computes the ratio of relevant predictions made by the classifier compared to all existing relevant instances that should have been retrieved. Recall evaluates the classifier capability of predicting the complete universe of relevant instances.

F-measure is an harmonic mean between Precision and Recall scores. F-β score is a generalization of the F-measure, that can introduce more weight on either Precision or Recall scores. Since the evaluation focus is on the ability of identifying all relevant instances, Recall is emphasized by using a β value greater than 1. In our experiments, we will apply $\beta = 2$, leading to the F-2 score. Matthews Correlation Coefficient computes a coefficient of agreement between the observed and predicted classifications.

4 Future Work

This project aims to develop a model able to handle the automatic text classification of HIV-related documents. This model will support the triage task in the curation process of HIV research. After designing and implenting a classification model, the system to be developed will be capable of selecting potentially relevant documents. To design the model, the experiments will be based in a supervised learning approach. Our goal is to experiment with several feature settings, machine learning algorithms and under-sampling factors to determine the most fitting configuration to tackle the HIV document triage, with regards to the dataset balance and computational costs.

References

[1] E. W. Sayers *et al.*, "Database resources of the national center for biotechnology information," *Nucleic acids research*, vol. 39, no. suppl 1, pp. D38–D51, 2011.

[2] D. Howe *et al.*, "Big data: The future of biocuration," *Nature*, vol. 455, no. 7209, pp. 47–50, 2008.

[3] T. B. Murdoch and A. S. Detsky, "The inevitable application of big data to health care," *JAMA*, vol. 309, no. 13, pp. 1351–1352, 2013.

[4] "Fourth biocreative challenge evaluation workshop."

[5] F. Leitner *et al.*, "Introducing meta-services for biomedical information extraction," *Genome Biol*, vol. 9, no. Suppl 2, p. S6, 2008.

[6] L. Hirschman *et al.*, "Text mining for the biocuration workflow," *Database*, vol. 2012, p. bas020, 2012.

[7] R. T.-H. Tsai, H.-J. Dai, P.-T. Lai, and C.-H. Huang, "Pubmed-ex: a web browser extension to enhance pubmed search with text mining features," *Bioinformatics*, vol. 25, no. 22, pp. 3031–3032, 2009.

[8] E. Pafilis, S. I. O'Donoghue, L. J. Jensen, H. Horn, M. Kuhn, N. P. Brown, and R. Schneider, "Reflect: augmented browsing for the life scientist," *Nature biotechnology*, vol. 27, no. 6, pp. 508–510, 2009.

[9] F. Boudin, J.-Y. Nie, and M. Dawes, "Clinical information retrieval using document and pico structure," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 822–830.

[10] C.-H. Wei, B. R. Harris, D. Li, T. Z. Berardini, E. Huala, H.-Y. Kao, and Z. Lu, "Accelerating literature curation with text-mining tools: a case study of using pubtator to curate genes in pubmed abstracts," *Database*, vol. 2012, p. bas041, 2012.

[11] H.-M. Müller, E. E. Kenny, and P. W. Sternberg, "Textpresso: an ontology-based information retrieval and extraction system for biological literature," *PLoS biology*, vol. 2, no. 11, p. e309, 2004.

[12] M. Miwa, J. Thomas, A. O'Mara-Eves, and S. Ananiadou, "Reducing systematic review workload through certainty-based screening," *Journal of biomedical informatics*, vol. 51, pp. 242–253, 2014.

[13] Y. Liu, N. V. Chawla, M. P. Harper, E. Shriberg, and A. Stolcke, "A study in machine learning from imbalanced data for sentence boundary detection in speech," *Computer Speech & Language*, vol. 20, no. 4, pp. 468–494, 2006.

[14] M.-L. Antonie, O. R. Zaiane, and A. Coman, "Application of Data Mining Techniques for Medical Image Classification." *MDM/KDD*, vol. 2001, pp. 94–101, 2001.

[15] G. Cohen, M. Hilario, H. Sax, S. Hugonnet, and A. Geissbuhler, "Learning from imbalanced data in surveillance of nosocomial infection," *Artificial Intelligence in Medicine*, vol. 37, no. 1, pp. 7–18, 2006.

[16] T. Fawcett and F. Provost, "Adaptive fraud detection," *Data mining and knowledge discovery*, vol. 1, no. 3, pp. 291–316, 1997.

[17] R. J. Bolton and D. J. Hand, "Statistical fraud detection: A review," *Statistical Science*, pp. 235–249, 2002.

[18] M. Kubat, S. Matwin *et al.*, "Addressing the curse of imbalanced training sets: one-sided selection," in *ICML*, vol. 97. Nashville, USA, 1997, pp. 179–186.

[19] R. Barandela, J. S. Sánchez, V. Garcıa, and E. Rangel, "Strategies for learning in class imbalance problems," *Pattern Recognition*, vol. 36, no. 3, pp. 849–851, 2003.

[20] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, "On the class imbalance problem," in *Fourth International Conference on Natural Computation, 2008. ICNC'08*, vol. 4. IEEE, 2008, pp. 192–201.

[21] P. Soda, "A multi-objective optimisation approach for class imbalance learning," *Pattern Recognition*, vol. 44, no. 8, pp. 1801–1810, 2011.

[22] N. Garca-Pedrajas, J. Prez-Rodrguez, M. Garca-Pedrajas, D. Ortiz-Boyer, and Colin, "Class imbalance methods for translation initiation site recognition in {DNA} sequences," *Knowledge-Based Systems*, vol. 25, no. 1, pp. 22 – 34, 2012, special Issue on New Trends in Data Mining.

[23] M. A. Maloof, "Learning when data sets are imbalanced and when costs are unequal and unknown," in *ICML-2003 workshop on learning from imbalanced data sets II, Washington DC*, vol. 2, 2003.

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority oversampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 341–378, 2002.

[25] 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*, 2007, pp. 35–41.

[26] L. Borrajo, R. Romero, E. L. Iglesias, and C. R. Marey, "Improving imbalanced scientific text classification using sampling strategies and dictionaries," *Journal of integrative bioinformatics*, vol. 8, p. 176, 2011.

[27] C. E. Lipscomb, "Medical subject headings (mesh)," *Bulletin of the Medical Library Association*, vol. 88, no. 3, p. 265, 2000.

[28] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," *Machine Learning*, vol. 59, no. 1-2, pp. 161–205, 2005.

[29] E. Charton, M. Meurs, L. Jean-Louis, and M. Gagnon, "Using collaborative tagging for text classification," *Informatics 2014*, pp. 32–51, 2013.

[30] V. N. Vapnik, "The Nature of Statistical Learning Theory," 1995.

[31] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Machine Learning: ECML 2004*. Springer, 2004, pp. 39–50.

[32] Y. Tang, Y.-Q. Zhang, N. V. Chawla, and S. Krasser, "Svms modeling for highly imbalanced classification," *Systems, Man, and Cybernetics, Part B: IEEE Transactions on Cybernetics*, vol. 39, no. 1, pp. 281–288, 2009.

[33] A. Mountassir, H. Benbrahim, and I. Berrada, "An empirical study to address the problem of unbalanced data sets in sentiment classification," *IEEE Systems, Man, Cybernetics*, pp. 3298–3303, 2012.