# Biomedical Literature Triage using Supervised Learning

**Hayda Almeida**[1], **Marie-Jean Meurs**[2,*], **Leila Kosseim**[1], **Greg Butler**[1,2], **Adrian Tsang**[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,leila.kosseim,gregory.butler,adrian.tsang}@concordia.ca

## Abstract

We present the mycoSORT system, a supervised machine learning based system to support the automatic triage of biomedical literature. This work reports on a total of 108 experiments combining 4 feature settings, 3 machine learning algorithms, and 9 under-sampling factors in a highly imbalanced corpus. The results show that the best approach relies on a classification model composed by domain annotations, a balanced dataset, and the use of a Logistic Model Trees classifier.

## 1 Introduction

Because genomics-based research of fungal enzymes benefits numerous industrial processes, substantial effort is devoted to the curation of the ever growing literature related to these enzymes. However, biomedical databases have recently experiencied a significant growth of available resources [1], and this phenomenon is expected to continue. Hence, the literature triage of such large databases represents a severe bottleneck in the manual curation workflow, since curators have to filter a massive list of documents, and select very few potential candidates to pass through the full curation process. In the context of our task, biocurators are seeking reference articles related to characterized lignocellulose-active proteins of fungal origin, in order to populate the mycoCLAP database [http://mycoclap.fungalgenomics.ca] [2]. mycoCLAP contains manually curated fungal enzymes with their reference articles. To create this database, biocurators had to verify an extensive amount of published literature in order to select, on average, only 10% of curatable documents among all search results. This low percentage indicates an imbalanced class distribution in the dataset, where the great majority of documents are classified as non-curatable, and only a small minority are classified as curatable.
In this work, we present a supervised machine learning approach to perform text classification of PubMed [3] abstracts, with the goal of supporting the triage of biomedical documents.

## 2 Related work

Imbalanced datasets are typical in bioinformatics and big data related tasks, where the relevant information usually represents a needle in a haystack. Other common real world situations that often deal with imbalance data include fraud and image detection [4][5], medical diagnosis [6][7] and speech recognition [8].
The imbalance issue has a great negative impact in the classifier performance. Being more represented in the dataset, the majority class can bias the classifier since more information would have been learned from majority instances than from minority instances.

*corresponding author

| | | | |
|---|---|---|---|
| Negative instances | 6,834 (90.12%) | # of words in paper abstracts | 43,598 |
| Positive instances | 749 (9.88%) | # of words in paper titles | 12,388 |
| Total # of instances | 7,583 (100%) | # of annotations in paper abstracts | 50,866 |
| Total # of abstracts with text content | 6,898 (90.96%) | # of annotations in paper titles | 8,172 |
| | | # of EC numbers | 12,272 |

Table 1: Statistics on the *mycoSet*

Different approaches have been proposed to deal with the imbalance data issue. Two well accepted methods are cost-sensitive classifiers, that are applied at the algorithm level, and data-sampling, applied at the data level. The main goal of the cost-sensitive approach [9] is to reduce the classification errors made in the minority class by introducing a bias (weight) in the classifier, so that classification mistakes on the minority class are more costly than in the majority class. Two data-sampling techniques were described by [10] in the application of the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE suggests a combination of under-sampling (i.e. reducing the majority class) and over-sampling (i.e. generating synthetic examples of the minority class), as an attempt to handle the imbalance in the dataset class distribution.

However, a study conducted by [11] evaluated both cost-sensitive and data-sampling approaches, and was not conclusive about the best option to tackle the data imbalance issue. Even though data-sampling has not outperformed different methods with regards to the imbalance problem, as shown by [9] and [12], when compared to other methods such as cost-sensitive, sampling is understood to be a less restrictive technique [11], therefore more easily applicable across different tasks. The under-sampling technique, an approach that reduces the number of majority instances until a specific percentage, is recommended by [11] as an option to cut down training time or to make a training phase feasible, in case the training dataset requires more computational resources than what is available to process it. In this paper, we analyze the use of under-sampling techniques with different classification algorithms and feature settings, implemented through the mycoSORT system. In Section 3, we describe how these techniques were applied in our dataset.

## 3 Methodology

**Corpus and Data Sampling.** The dataset applied in our experiments is composed of journal paper abstracts retrieved from PubMed by biocurators. Queries used to retrieve these documents were formed by a *name of an enzyme (family) of interest*, the logical conjunction *AND*, and the generic string *fung** to match fungal-related terms. The abstracts used to form our dataset were publish before December 31, 2013. We call this corpus *mycoSet*. *mycoSet* was also preprocessed with the mycoMINE text mining system [13], which added bio-entity annotations to relevant units of text. All document instances in *mycoSet* were correctly labeled by biocurators with a positive (curatable) or negative (non-curatable) class. With the manual labelling effort, biocurators classified 749 as positive instances and rejected 6,834 instances. *mycoSet* contains 7,583 document instances and is highly imbalanced, as shown in table 1, which presents further details about the corpus. Thus, for creating our training set, we applied the under-sampling technique, with the goal of dealing with both the imbalance issue and the large dataset size. The generation of training sets started with a class distribution that represents the real imbalanced scenario of the task (90% of negative instances and 10% of positive instances). Subsequently, an under-sampling factor (USF) of 5% was applied, to reduce the majority class representation, until the class distribution was similar for both positive and negative (50% each).

**Features.** Relevant fragments of text were extracted from the documents to be used as features in our classification models. Features were mostly selected from the "AbstractText", "ArticleTitle" and "RegistryNumber" text fields. Bioentities, annotated with mycoMINE [13], were extracted from these fields and grouped by their span: sentence or entity level. As the sentence level annotation corresponds to an entire sentence, these features were represented using a bag-of-words approach. Differently, features of the entity level were kept as they appear in the document.

Each feature vector is then built to represent a document by accounting for the occurrence of features in its text, and its classification label. The large dataset size of this task results in a large and sparse

matrix representation, leading to a costly computational processing. We studied techniques to reduce the feature space size by applying standard feature selection methods, such as performing filtering by character length and occurrence. The filters were applied in a way such that words occurring only once in the training corpus or that contain no more than 3 characters were discarded when generating feature vectors.

**Classification Algorithms.** In our experiments, we made use of three different classification algorithms: Naïve Bayes (NB), Logistic Model Trees (LMT) and Support Vector Machine (SVM).

NB is used as a baseline evaluation of our sampling and feature selection strategies. NB assumes a strong conditional independence of features, considering that in a feature vector $F$, the features $f_1, ..., f_n$ are conditionally independent from each other, given a class $C$. LMT [14] was previously described by [15] as a classifier that is able to efficiently handle tasks with imbalanced datasets. It consist of a combination of Decision Tree and LogitBoost algorithms, being a classification tree, with logistic regression models on its nodes. SVM [16] was also recommended by previous works as an algorithm able to deal with imbalanced data [17, 18, 19]. SVM computes the "margin maximum classifier" [20], the largest radius around a classification boundary, and tries to separate data points on a dimensional space, and to identify the different classes to which they belong.

**Evaluation Metrics.** Our experimental results are evaluated in terms of Precision, Recall, F-measure, and F-2. While F-measure is the harmonic mean between Precision and Recall, F-$\beta$ is a generalization of the F-measure, that can favor either Precision or Recall. Since we focus on evaluating the model capability of identifying the entire universe of relevant instances, we emphasize Recall by using a $\beta$ value greater than 1. In our experiments, presented in Section 4, we applied $\beta$ = 2, leading to the F-2 score.

## 4 Experiments and Results

**Experiments.** We experimented with 108 different classification models, created from a variation of the following variables (4 set of features, 3 classifiers and 9 Under-Sampling Factors) combined in different configuration settings. The groups of features used across experiments were defined as follows: [F1: Annotated bio-entities], [F2: Annotated contents of entity spans], [F3: Annotated contents of sentence spans (as a bag-of-words)], [F4: Enzyme Commission (EC) numbers], [F5: Bag-of-words representation of the entire fields (ArticleTitle and AbstractText)].

The system tested three classifiers: Naïve Bayes (NB), Logistic Model Trees (LMT), and Support Vector Machine (SVM). As described in Section 3, USF of 5% was applied until the class distribution was similar for both positive and negative.

The classification models were separated by 5 set of experiments. All sets of experiments were evaluated across all classifiers and USFs. The set of experiments S1 is formed by only 22 Bio-entities [F1] as features. The set of experiments S2 is composed by the 22 bio-entities [F1] plus the EC numbers [F4] listed in the training set. The features in the set of experiments S3 are the bag-of-words representation of the text fields [F5]. Features in the set of experiments S4 are a combination of the previous sets: the 22 bio-entities [F1], their annotated content [F2, F3] and the EC Numbers [F4] listed in the training set.

**Results.** The summary of F-measure and F-2 scores for all the experiments can be seen in Figure 1. The set of experiments S3 is considered as our baseline, since features are only represented as a bag of words. When comparing the results of S1 and S2 to this baseline, we notice that the bag-of-words approach (S3) shows better classification scores. This difference is explained by the size of the feature space, that can influence the classification model building. On one hand, S1 has only 22 features and S2 has from 186 (in the most balanced training set) to 397 features (in the most imbalanced training set); on the other hand, S3 has from 7,622 to 20,729 features. A much larger feature space provides more information for building the classification models, but it also brings a data sparseness problem, requiring extra training time and computational resources.

A comparison between the results from S1 and S2 show a better classification performance when the EC numbers are taken into account in the set of features. However, the classification results obtained from S1 illustrate an interesting cost-benefit, since this set uses a very restricted set of features and is still capable of achieving reasonable scores compared to the other more robust models.

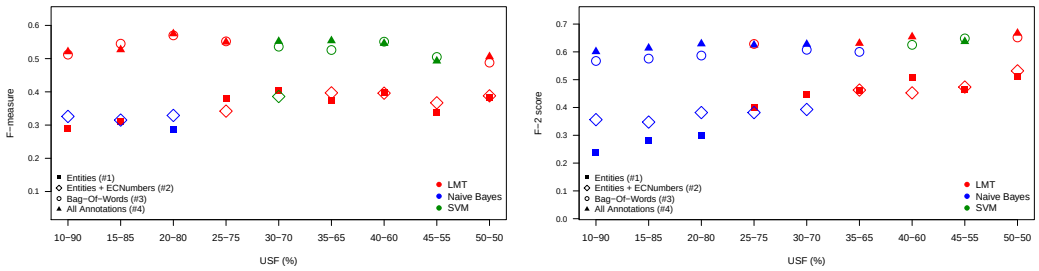


Figure 1: F-measure and F-2 scores of best classifiers for each model

Another interesting comparison is between S3 and S4; Even though the feature space in S4 is smaller, with 3,338 to 8,931 features, it still demonstrates better results than S3. Therefore, an improved performance is noticed when using domain annotations as features.

The classification scores demonstrated that the under-sampling strategy also supported the performance enhancement, as observed by [12, 19], independently of the feature set applied. Better results were achieved by the experiments that used a training corpus with a more balanced class distribution, even though the test set presented a realistic and imbalanced class distribution ratio.

In our experiments, although the NB classifier was considered as a baseline, it was still able to outperform all other classifiers for highly imbalanced class distribution with small feature spaces. That could possibly indicate that the NB classifier is less sensitive to the class distribution in the dataset. The SVM classifier demonstrated better performance for cases in which the USF varied from 20% to 35% and the feature space was larger, as with the use of the bag-of-words approach. LMT was able to outperform the other classifiers in models characterized by less class imbalance and a very restricted feature space. The LMT performance under a feature space of size 22 is comparable to the classification performance of a model that is composed by 186 features. This observation comfirmed results before described by [15] about the LMT ability of handling harder classification tasks.

## 5 Conclusion

This study presents the mycoSORT system that relies on a supervised machine learning approach to perform automatic text classification of PubMed abstracts. The objective of mycoSORT is to support and improve the triage of candidate articles for the mycoCLAP database. The mycoSORT system learns from correctly classified samples of candidate PubMed abstracts, then provides a classification decision for a new document according to specific attributes found in the data.

We experimented with 4 feature settings, 3 machine learning algorithms, and 9 under-sampling factors, for a total of 108 experiments that allowed to analyze the discriminative power of different data attributes to best predict the document relevance, and study the effect of different sampling techniques to overcome the dataset imbalanced characteristic.

The results demonstrate that in our context, the best approach to deal with the triage of imbalanced corpora relies on a classification model composed by domain annotations, a balanced dataset, and the use of a LMT classifier. Moreover, the other models studied here can be used as further options to the automatic document classification in the triage task, in case of existing constraints related to computational cost or data availability.

The mycoSORT system is fully implemented, and publicly released as an open source toolkit available at `https://github.com/TsangLab/mycoSORT`. The *mycoSet* corpus used in our experiments is also publicly available as a list of pairs [abstract PubMed ID - class of the abstract], making all our experiments fully reproducible.

## References


[1] L. Hunter and K. B. Cohen, "Biomedical language processing: Perspective what's beyond PubMed?" *Molecular cell*, vol. 21, no. 5, p. 589, 2006.

[2] C. Murphy *et al.*, "Curation of characterized glycoside hydrolases of fungal origin," *Database*, 2011.

[3] National Center for Biotechnology Information, "PubMed," Bethesda, MD, USA, 2005.

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

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

[6] 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.

[7] G. Cohen *et al.*, "Learning from imbalanced data in surveillance of nosocomial infection," *Artificial Intelligence in Medicine*, vol. 37, no. 1, pp. 7–18, 2006.

[8] Y. Liu *et al.*, "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.

[9] 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.

[10] 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.

[11] 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*, 2007, pp. 35–41.

[12] L. Borrajo, R. Romero, E. L. Iglesias, and C. R. Marey, "Improving imbalanced scientific text classification using sampling strategies and dictionaries," *J. of Integrative Bioinformatics*, vol. 8, p. 176, 2011.

[13] M. Meurs *et al.*, "Semantic text mining support for lignocellulose research," *BMC Medical Informatics and Decision Making*, vol. 12, 2012.

[14] N. Landwehr, M. Hall, and E. Frank, "Logistic Model Trees," *Machine Learning*, vol. 59, 2005.

[15] E. Charton, M. Meurs, L. Jean-Louis, and M. Gagnon, "Using Collaborative Tagging for Text Classification," *Informatics 2014*, pp. 32–51, 2013.

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

[17] R. Akbani, S. Kwek, and N. Japkowicz, "Applying Support Vector Machines to Imbalanced Datasets," in *ECML 2004*. Springer, 2004, pp. 39–50.

[18] Y. Tang, Y.-Q. Zhang, N. V. Chawla, and S. Krasser, "SVMs modeling for highly imbalanced classification," *IEEE Systems, Man, and Cybernetics*, pp. 281–288, 2009.

[19] A. Mountassir, H. Benbrahim, and I. Berrada, "An Empirical Study to Address the Problem of Unbalanced Data Sets in Sentiment Classification," *IEEE Systems, Man, and Cybernetics*, pp. 3298–3303, 2012.

[20] S. Marsland, *Machine Learning: An Algorithm Perspective*. Chapman and Hall, 2009.