

# Multilabel text classification of unbalanced datasets: Two-pass NNMF

Gabriella Skitalinskaya<sup>1</sup> and John Cardiff<sup>1</sup>

<sup>1</sup>Institute of Technology, Tallaght, Dublin  
gabriellasky@icloud.com; john.cardiff@itttdublin.ie

**Abstract.** The natural distribution of textual data used in text classification is often imbalanced. Categories with fewer examples are underrepresented and their classifiers trained on the datasets transformed to bag-of-words representations or basic topic modeling transformations often perform far below a satisfactory level. We tackle this problem using a two-pass non-negative matrix factorization algorithm. This approach finds topics for each category independently allowing to better define topics for underrepresented categories. The results are analyzed from multiple goal perspectives - H-loss, accuracy, F-measure, precision, and recall, from the micro, macro and example-based aspect since each is appropriate in different situations. Through experimental validation, it is shown that the two-pass matrix factorization improves classification results achieved using bag-of-words representations.

**Keywords:** topic modeling, matrix decomposition, multi-label text classification

## 1 Introduction

### 1.1 Problem setting

Multi-label classification is a predictive data mining problem which is applicable to a wide variety multiple real-world problems, including the automatic labeling of many resources such as texts, images, music, and video [1–5].

One of the most popular problems in this domain is text categorization (text classification), organizing text documents into several not mutually exclusive categories. The algorithms used for multi-label classification can be grouped into two classes: discriminative algorithms and generative modeling algorithms. The discriminative algorithms extend single-label algorithms so they can handle multi-label data. Excellent reviews with comparisons of discriminative algorithms are presented in [6, 7]. The generative modeling algorithms model multi-label collections via the Bayes rule. According to [8–10] supervised topic models have become one of the leading generative modeling algorithms. Both classes of algorithms have their own disadvantages. The discriminative algorithm is often prone to over-fitting and highly skewed datasets, while the generative modeling algorithm may ignore some obvious observed features. Combining two classes of algorithms allows to pursue more robust algorithms.

Instead of applying multi-label classification algorithms directly to the bag-of-words representations of document collections, in this paper we propose a two pass matrix decomposition approach based on non-negative matrix factorization (NNMF), which captures topics in a corpus of documents. The algorithm was introduced in [11] in the context of dynamic topic modeling. The two-pass matrix decomposition approach is an unsupervised technique for topic modeling, that can automatically identify topics for each category/label independently, thus, it is able to identify topics even from underrepresented categories. Representing texts according to their topic distributions is more compact than bag-of-words representation and can be processed faster than raw text in subsequent automated processes.

By using the transformed topic mixture proportions as a new representation of documents, we obtain an unsupervised dimensionality reduction algorithm that uncovers the latent structure in a document collection while preserving predictive power for the problem of classification. We demonstrate the proposed approaches applicability by analyzing news articles(Reuters) and scientific article abstracts(BibTex).

The paper is organized as follows. Section 2 describes the proposed two-pass algorithm. Measures for evaluation of topic models and classification are described in Section 3. Section 4 provides information on the dataset used for topic modeling. Information on the experimental setup, base classifiers parameters and results of the application of the proposed algorithm to the dataset are presented in Section 5. Section 6 provides a comparison of results obtained by different classifiers using data represented as topic distributions obtained by the two-pass approach with the bag-of-words data representations. Finally, Section 7 presents our conclusions.

## 1.2 Related Work

Two main approaches for solving multi-label classification problems can be identified: problem transformation methods and algorithm adaptation methods. The former transforms the multi-label problem into a single-label multi-class problem that is solved with single-label classification algorithms, e.g., binary relevance[12], classifier chains [13], random k-labelsets [14], and conditional dependency network [15], whereas the latter consists of extending a single-label algorithm so it can handle multi-label data, e.g., rank support vector machines (SVMs) [16], multi-label C4.5 [17], multi-label k nearest neighbors [18], multilabel neural networks [19], CLR [4], HOMER [20] and ECC [21].

Learning from imbalanced data is a problem which arises in many real-world datasets. Much progress has been made in developing learning algorithms dealing with imbalance based on algorithmic adaptations [22, 23], the use of ensembles [24] and resampling techniques [25–27]. One of the most deeply studied approaches lately is dealing with imbalance using resampling methods. Among the existing resampling techniques, those based on the creation of new samples (oversampling) have shown to work better than others [28]. The new samples can be clones of existent ones, or be synthetically produced as in MLSMOTE

(MultiLabel Synthetic Minority Over-sampling Technique) [29]. Multilabel over-sampling algorithms based on the cloning approach proposed in [25, 26] demonstrate the approaches capability to improve classification results. In [27] a new multi-label learning approach named cross-coupling aggregation (COCOA) is proposed which is aimed at leveraging the exploitation of label correlations as well as the exploration of class-imbalance.

We have selected the following well-known and widely used methods from the literature for our benchmark comparison: Binary Relevance, Classifier Chains, ML-kNN, RAkEL and propose our own approach to dealing with unbalanced datasets.

## 2 Approach

### 2.1 Background

Non-negative matrix factorization (NNMF) is a matrix decomposition approach which decomposes a non-negative matrix into two low-rank non-negative matrices [30]. The main difference between NNMF and other classical matrix decomposition methods relies on the non-negativity constraints imposed on the model. These constraints tend to lead to a parts-based representation of the data because they allow only additive, not subtractive, combinations of data items. In this way, the factors produced by this method can be interpreted as parts of the data or, in other words, as subsets of elements that tend to occur together in sub-portions of the dataset.

Formally, non-negative matrix decomposition can be described as:  $V \approx WH$  where  $V \in R^{m \times n}$  is a positive data matrix with  $m$  variables and  $n$  objects,  $W \in R^{m \times k}$  are the reduced  $k$  basis vectors or factors, and  $H \in R^{k \times n}$  contains the coefficients of the linear combinations of the basis vectors needed to reconstruct the original data.

In the context of text analysis, for example, matrix  $V$  can be represented as a Document-Term matrix, where  $m$  is the number of documents and  $n$  the number features, matrix  $W$  represents a Document-Topic matrix, and matrix  $H$  - the Topic-Term matrix.

### 2.2 Two-pass topic modeling algorithm

First of all, for the purposes of the present paper the following definitions are used. The entire collection is divided into subsets of documents each containing a specific label, hereinafter "label subset". The label subsets may overlap, since each document may have several labels. Each document may reflect one or more topics. Each topic is represented by its top-terms. Top-terms are terms that have the highest frequency (on average) in those documents that contain the topic. The number of top-terms for all topics, regardless of the category label, is the same and is assigned by the user (for example, 5, 10, 20, etc.). When applying matrix decompositions to each label subset the user must specify the number of

topics. One of the quality measures that allows us to choose the best number of topics is the so-called coherence measure [31–34].

When applying topic modeling to the entire collection, the algorithms prove to be insensitive to the topics reflected only by a small fraction of the documents, which is the typical for multi-label classification tasks. The main hypothesis of our approach is that independently modeling topics for subsets of documents for each label and subsequently aggregating the found topics into one matrix allows capturing the underrepresented topics in the collection. In the case of multi-label classification, due to the fact that subsets of documents for each label may overlap, reapplication of the matrix decomposition to the aggregated matrix allows to combine topics that are common for several labels. This reapplication of the matrix decomposition reduces the dimensionality of the data, which allows to reduce computation costs.

The approach is represented by the following algorithm:

First pass. NNMF is applied to each label subset. As a result, for each label a set of  $k$  topics is obtained, where  $k$  is defined by the user. Topics are described by a user-specified number of top-terms  $t$  and a set of all related documents.

Data Transformation. Using the topic models obtained after the first pass we construct a new compressed representation, looking through the rows of each Topic-Term matrix of each label topic model. Each row contains weights of all the terms of a particular topic of the label topic model under consideration. We construct the new Topic-Term matrix with two subsequent procedures:

- (1) In each topic from each label topic model, the top- $t$  terms are taken from the appropriate topic-term matrix, all weights for the remaining terms are set to 0.
- (2) The obtained vectors for all label topic models are combined into one matrix.

Second pass. NNMF is re-applied to the transformed data, outputting a set of more general topics, each of which has a set of label topics associated with it. By applying matrix decompositions in this step, we identify  $k'$  general topics that potentially include topics from several labels. The number of general topics  $k'$  to be found in this step is specified by the user.

The matrix has the size  $m \times n$ , where  $m$  is the total number of topics in all label models, and  $n$  is the subset of the terms remaining after the data transformation. By using only the top- $t$  terms in each topic we include only the terms that were important in any label and exclude the terms that never figured in any label topic.

### 3 Quality of classification

#### 3.1 Measures for evaluation of topic modeling

Coherence measures evaluate the interpretability of the automatically generated topics and find the best number of topics. The higher the coherence score, the

better the topic model. The most widely used coherence measures to determine the optimal number of topics in each time window and the optimal number of dynamic topics, such as *UCI* [31], *NPMI* [32], *C<sub>v</sub>* [33]. But according to [34] the TC-W2V [34] measure outperforms them.

The TC-W2V score uses the widely known word2vec tool [35] to create term vectors. In this paper we have used the Skip-gram algorithm, which predicts context words based on the current word, for estimating word representations in a vector space. The coherence of a topic represented by its  $t$ -ranked terms is determined by the mean pairwise cosine similarity between  $t$  corresponding vector-terms in the word2vec space:

$$coh(t_h) = \frac{1}{\binom{t}{2}} \sum_{j=2}^t \sum_{i=1}^{j-1} \cos(wv_i, wv_j).$$

A general evaluation of the coherence of a topic model  $T$ , consisting of  $k$  topics, is determined by the mean of individual topic coherence scores:

$$coh(T) = \frac{1}{k} \sum_{h=1}^k coh(t_h).$$

### 3.2 Measures for evaluation of classification

Performance evaluation for multi-label learning systems differs from that of classical single-label learning systems. For example, a prediction could be partially correct (some of the labels are correctly predicted), fully correct (all label predictions are correct), or fully incorrect (predictions for all labels are wrong). It is essential to include multiple and contrasting measures because of nature of the multi-label classification setting.

Metrics to evaluate bipartitions can be classified into two groups: label-based and example-based. The example-based evaluation measures are based on the average differences of the actual and the predicted sets of labels over all examples of the evaluation dataset. The label-based evaluation measures, on the other hand, assess the predictive performance for each label separately and then average the performance over all labels.

Two different label-based approaches can be used: macro and micro. Micro averaged scores give equal weight to every example and tend to be dominated by the performance in most common categories. Macro averaged scores give equal weight to every category, regardless of its frequency and is more influenced by the performance on rare categories. The macro approach is used when the system is required to perform consistently across all classes regardless of the frequency of the class (i.e., in problems where distribution of training samples across categories is skewed), whereas the micro approach may be better if the density of the class is important.

In our experiments, we used five example-based evaluation measures (Hamming loss, accuracy, precision, recall, F1 score) and six label-based evaluation

measures (micro-precision, micro-recall, micro-F1, macro-precision, macro-recall and macro-F1).

## 4 Dataset Description

Since we are interested in evaluating the strengths and weaknesses of our approach for different classification algorithms in a multi-label text classification context, we decided to use datasets with diverse characteristics. In our experiments to ensure comparability with other works we used two popular benchmark datasets (Reuters, BibTex) with textual data for evaluation and comparison. These datasets come pre-divided into training and testing parts: thus, in the experiments, we use them in their original format.

Reuters-21578 is one of the most widely used benchmarking collection for text categorization problems [36]. The corpus consists of news articles that appeared in the Reuters newswire in 1987. The BibTex dataset<sup>1</sup> is a large collection of scientific article abstracts tagged by users using 159 tags.

The obtained datasets have varying feature to label ratios and cardinality.

The key statistics for the mentioned datasets are presented in Table 1. Every document from the dataset collections went through the following preprocessing procedures:

- removal of stopwords
- removal of short words (less than 3 characters)
- lemmatization.

Also TF-IDF term weighting and document length normalization is applied to the Document-Term matrices for each label subset.

Table 1: Dataset key characteristics

Dataset	Labels	Training	Test	Features	Cardinality	Average num. of words per document
<i>Reuters</i>	108(55)	7713	2987	8859	1.22	162
<i>BibTex</i>	159	4880	2515	1836	2.40	60

## 5 Experiments

### 5.1 Experimental setup

Word2vec models and non-negative matrix factorization has been carried out using the gensim<sup>2</sup> and sklearn<sup>3</sup> Python libraries. The comparison of the multi-label learning methods was performed using the implementations in the following

<sup>1</sup> The dataset can be downloaded at: <http://mulan.sourceforge.net/datasets.html>

<sup>2</sup> <https://radimrehurek.com/gensim/>

<sup>3</sup> <http://scikit-learn.org/>

library: scikit-multilearn. Scikit-multilearn is a BSD-licensed library for multi-label classification that is built on top of the well-known scikit-learn ecosystem.

## 5.2 Base Classifiers

To ensure comparability with other works we used the same parameter values recommended by the authors of the following papers [13, 6] or by the authors of other relevant publications.

The methods used in the mentioned papers use SVMs as base classifiers for solving the partial binary classification problems in all problem transformation methods and the ensemble methods.

In particular, [6] used the implementation based on libsvm for training SVMs with a linear basis, but in [6] the authors trained SVMs with a radial basis kernel for all problem transformation methods and RAKEL. The kernel parameter gamma and the penalty C, for each combination of dataset and method, are determined by using 10-fold cross validation only on training sets. As proposed by the authors the values  $2^{-15}, 2^{-13}, \dots, 2^1, 2^3$  were considered for gamma and  $2^{-5}, 2^{-3}, \dots, 2^{13}, 2^{15}$  for the penalty C.

The number of neighbors in the ML-kNN method for each dataset is determined from the values 6 to 20 with step 2.

The number of models in RAKEL is set to  $\min(2Q, 100)$ , where Q is the number of labels for all datasets, the size of the label-sets  $k$  is set to half the number of labels ( $Q/2$ ) [14]. The ensemble iterations (where relevant) are set to  $m = 50$ .

The best parameters are determined for every method on each dataset.

## 5.3 Topic Modeling

To find the optimal number of topics for each label and the optimal number of general topics for each dataset using the two pass algorithm a user-specified number of top-terms used to determine the coherence of obtained topics is needed.

Due to the fact the average number of words per document in the Reuters dataset is quite high, the top-terms parameter has been set to  $n = 10$  words. For the BibTex dataset, where the average number of words per document is significantly lower, the number of top-terms chosen for finding the optimal number of topics per label has been set to  $n = 5$ .

Depending on the number of top-terms after the first pass of the proposed approach we obtain an aggregated Topic-Term matrix for each dataset, which can be used for reducing the dimensionality of the initial dataset. Finding topics for each label independently results in overlapping topics for two or more labels. During the second pass of the proposed approach such topics are combined into more general topics, reducing the dimensionality of the data even further. Thus, for the Reuters dataset we have reduced the number of initial features from 8859 to 100, and from 1836 to 280 for the BibTex dataset. The number of features after each pass of the proposed algorithm is shown in Table 2

Table 2: Number of features after each step of the two-pass approach

	<b>Reuters</b>	<b>BibTex</b>
<b>Initial Num. of Features</b>	8859	1836
<b>First Pass.</b>		
<i>Num. of Aggregated Topics</i>	323	1088
<b>Second Pass.</b>		
<i>Num. of General Topics</i>	100	280

After obtaining the general topics vectors found in the document collection, the training set is transformed using non-negative matrix factorization, solving the following problem: Given a non-negative matrix, find non-negative matrix factors  $W$  and  $H$  such that:  $X \approx WH$ .

We apply NMF to the initial Document-Term matrix obtained for the training set using the precomputed General Topic- Term matrix found by the two-pass approach as  $H$ . As output we get  $W$  - the Document-General Topic matrix, which will be used as input for the classification task.

## 6 Classification Results

In this paper the classifiers are applied to data representations obtained after the two-pass NMF algorithm and compared to the baseline. As the baseline we have chosen classifiers trained on bag-of-words data representations (BL1 - [13], BL2 - [6]). Classifiers are evaluated by their performance when applied to the test set from the corresponding dataset.

Results in terms of H-Loss and example-based Accuracy, Precision, Recall and F-measure are shown in Table 3 with the appropriate baseline values where possible. Since in multi-label classification different evaluation measures are appropriate in different tasks it is expected that a method may not outperform others in all measures. Table 4 gives the precision, recall and F1 scores using micro averaging, while Table 5 gives the corresponding values obtained by macro averaging along with the appropriate baseline values where possible.

As for the Reuters dataset, one can see a significant increase in the values of H-loss and example-based measures for all methods when using the two-pass NMF approach. RAKEL performs best according to the four example-based evaluation measures, but BR performs best according to H-loss.

As for the BibTex dataset, the proposed approach performs better than the baseline only when using ML-kNN classifier according to all the used evaluation measures except for micro-precision. Though when comparing the evaluation measures obtained by ML-kNN with values obtained by other measures, it can be seen that the latter perform better.

It can be seen that the overall performance achieved on the Reuters dataset is higher than on the BibTex dataset. This could be explained by the fact that the texts in the Reuters dataset are longer and more suitable for topic modeling. Learning topics from short texts is considered to be a challenging problem due

to the severe sparsity of Document-Term data matrix, since the texts in BibTex dataset are very short(60 words per document on average), the obtained topics may be of low quality.

Table 3: Comparison of classification results using BoW and two-pass NNMF input transformation

Reuters								
	BR		CC		ML-KNN		Rakel	
	<i>BL1</i>	<i>2-Pass NNMF</i>	<i>BL1</i>	<i>2-Pass NNMF</i>	<i>BL1</i>	<i>2-Pass NNMF</i>	<i>BL1</i>	<i>2-Pass NNMF</i>
H-Loss	0.011	<b>0.008</b>	0.011	<b>0.009</b>	-	<b>0.009</b>	0.011	<b>0.009</b>
Accuracy	0.319	<b>0.727</b>	0.387	<b>0.753</b>	-	<b>0.728</b>	0.337	<b>0.785</b>
Precision	-	<b>0.807</b>	-	<b>0.820</b>	-	<b>0.792</b>	-	<b>0.855</b>
Recall	-	<b>0.811</b>	-	<b>0.811</b>	-	<b>0.781</b>	-	<b>0.841</b>
F1	0.222	<b>0.799</b>	0.250	<b>0.808</b>	-	<b>0.779</b>	0.233	<b>0.840</b>

  

BibTex								
	BR		CC		ML-KNN		Rakel	
	<i>BL2</i>	<i>2-Pass NNMF</i>	<i>BL2</i>	<i>2-Pass NNMF</i>	<i>BL2</i>	<i>2-Pass NNMF</i>	<i>BL2</i>	<i>2-Pass NNMF</i>
H-Loss	<b>0.012</b>	0.018	<b>0.012</b>	0.023	0.014	<b>0.014</b>	-	<b>0.017</b>
Accuracy	<b>0.194</b>	0.112	<b>0.202</b>	0.101	0.056	<b>0.080</b>	-	<b>0.126</b>
Precision	<b>0.515</b>	0.371	<b>0.508</b>	0.320	0.254	<b>0.277</b>	-	<b>0.372</b>
Recall	<b>0.373</b>	0.362	<b>0.378</b>	0.369	0.132	<b>0.170</b>	-	<b>0.299</b>
F1	<b>0.433</b>	0.344	<b>0.434</b>	0.316	0.174	<b>0.193</b>	-	<b>0.304</b>

Table 4: Comparing classification using BoW with two-pass NNMF input transformation micro P,R,F1

Reuters								
	BR		CC		ML-KNN		Rakel	
	<i>BL1</i>	<i>2-Pass NNMF</i>	<i>BL1</i>	<i>2-Pass NNMF</i>	<i>BL1</i>	<i>2-Pass NNMF</i>	<i>BL1</i>	<i>2-Pass NNMF</i>
Precision		0.863		0.821		0.835		0.828
Recall		0.751		0.745		0.711		0.775
F1		0.803		0.781		0.768		0.800

  

BibTex								
	BR		CC		ML-KNN		Rakel	
	<i>BL2</i>	<i>2-Pass NNMF</i>	<i>BL2</i>	<i>2-Pass NNMF</i>	<i>BL2</i>	<i>2-Pass NNMF</i>	<i>BL2</i>	<i>2-Pass NNMF</i>
Precision	<b>0.753</b>	0.39	<b>0.744</b>	0.291	<b>0.819</b>	0.585	-	<b>0.425</b>
Recall	0.328	0.328	0.335	<b>0.338</b>	0.118	<b>0.147</b>	-	<b>0.260</b>
F1	<b>0.457</b>	0.357	<b>0.462</b>	0.313	0.206	<b>0.235</b>	-	<b>0.323</b>

Table 5: Comparing classification using BoW with two-pass NNMF input transformation macro P,R,F1

Reuters								
	BR		CC		ML-KNN		Rakel	
	<i>BL1</i>	<i>2-Pass NNMF</i>	<i>BL1</i>	<i>2-Pass NNMF</i>	<i>BL1</i>	<i>2-Pass NNMF</i>	<i>BL1</i>	<i>2-Pass NNMF</i>
<b>Precision</b>	-	0.667	-	0.550	-	0.629	-	0.594
<b>Recall</b>	-	0.416	-	0.368	-	0.346	-	0.434
<b>F1</b>	-	0.474	-	0.420	-	0.416	-	0.477

  

BibTex								
	BR		CC		ML-KNN		Rakel	
	<i>BL2</i>	<i>2-Pass NNMF</i>	<i>BL2</i>	<i>2-Pass NNMF</i>	<i>BL2</i>	<i>2-Pass NNMF</i>	<i>BL2</i>	<i>2-Pass NNMF</i>
<b>Precision</b>	<b>0.528</b>	0.274	<b>0.539</b>	0.236	0.192	<b>0.240</b>	-	<b>0.265</b>
<b>Recall</b>	<b>0.250</b>	0.243	<b>0.257</b>	0.252	0.049	<b>0.071</b>	-	<b>0.169</b>
<b>F1</b>	<b>0.307</b>	0.229	<b>0.316</b>	0.222	0.065	<b>0.096</b>	-	<b>0.195</b>

## 7 Conclusions

In this paper, we present an algorithm based on two-pass NNMF for data transformation to improve classification results for multi-label learning with unbalanced classes. The proposed approach is able to identify topics even from under-represented categories. We evaluate the most popular methods for multi-label learning using a wide range of evaluation measures on two widely used benchmark datasets containing textual data. We compare our results to those obtained for bag-of-words representations. Our proposed topic based classifier system is shown to be competitive with existing text classification techniques.

Through experimental validation, it is shown that representing texts according to their topic distributions using the proposed two-pass approach improves classification results achieved using bag-of-words representations for longer texts such as news articles. As for shorter texts, such as abstracts for scientific articles, there is no significant increase in performance, but the achieved results are comparable to those obtained for bag-of-words data representations. This could be explained by the topic modeling nature of the proposed algorithm, as topic modeling of short texts is a problem yet to be tackled and simple matrix decomposition as NNMF used in this paper may not be able to obtain topics of high quality for short texts. Overall, representing texts according to their topic distributions is more compact than bag-of-words representation and can be processed faster than raw text in subsequent automated processes.

## References

1. Schapire, R.E., Singer, Y.: BoosTexter: A Boosting-based System for Text Categorization. Machine Learning **39** (2000) 135–168

2. Godbole, S., Sarawagi, S.: Discriminative Methods for Multi-labeled Classification. *Lecture Notes in Computer Science* **3056** (2004) 22–30
3. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. *Pattern Recognition* **37** (2004) 1757–1771
4. Fürnkranz, J., Hüllermeier, E., Loza Mencía, E., Brinker, K.: Multilabel classification via calibrated label ranking. *Machine Learning* **73** (2008) 133–153
5. Dimou, A., Tsoumakas, G., Mezaris, V., Kompatsiaris, I., Vlahavas, I.: An empirical study of multi-label learning methods for video annotation. In: 7th International Workshop on Content-Based Multimedia Indexing, CBMI 2009. (2009) 19–24
6. Madjarov, G., Kocev, D., Gjorgjevikj, D., Džeroski, S.: An extensive experimental comparison of methods for multi-label learning. In: *Pattern Recognition*. Volume 45. (2012) 3084–3104
7. Moyano, J.M., Gibaja, E.L., Cios, K.J., Ventura, S.: Review of ensembles of multi-label classifiers: Models, experimental study and prospects. *Information Fusion* **44** (2018) 33–45
8. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (2009) 248–256
9. Rubin, T.N., Chambers, A., Smyth, P., Steyvers, M.: Statistical topic models for multi-label document classification. *Machine Learning* **88** (2012) 157–208
10. Ma, H., Chen, E., Xu, L., Xiong, H.: Capturing correlations of multiple labels: A generative probabilistic model for multi-label learning. *Neurocomputing* **92** (2012) 116–123
11. Skitalinskaya, G., Alexandrov, M., Cardiff, J.: Comparison of two-pass algorithms for dynamic topic modeling based on matrix decompositions. In: *Proc. of 15th Mexican Intern. Conf. on Artif. Intell. (MICAI-2017)*, Springer, LNAI **10632-10633** (2017) 16 pp
12. Tsoumakas, G., Katakis, I., Overview, A.: Multi-Label Classification : An Overview. *International Journal of Data Warehousing and Mining* **3** (2007) 1–13
13. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. *Machine Learning* **85** (2011) 333–359
14. Tsoumakas, G., Katakis, I., Vlahavas, I.: Random k-labelsets for multilabel classification. *IEEE Transactions on Knowledge and Data Engineering* **23** (2011) 1079–1089
15. Guo, Y., Gu, S.: Multi-label classification using conditional dependency networks. *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Two* (2011) 1300–1305
16. Elisseeff, a.: Kernel methods for multi-labelled classification and categorical regression problems. *Advances in neural information processing* (2002) 1–18
17. Clare, A., King, R.D. In: *Knowledge Discovery in Multi-label Phenotype Data*. Springer Berlin Heidelberg, Berlin, Heidelberg (2001) 42–53
18. Zhang, M.L., Zhou, Z.H.: ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition* **40** (2007) 2038–2048
19. Zhang, M.L.: ML-rbf: RBF Neural Networks for Multi-Label Learning. *Neural Processing Letters* **29** (2009) 61–74
20. Tsoumakas, G., Katakis, I., Vlahavas, I.: Effective and efficient multilabel classification in domains with large number of labels. *Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD’08)* (2008) 30–44

21. Read, J., Pfahringer, B., Holmes, G.: Multi-label classification using ensembles of pruned sets. In: Proceedings - IEEE International Conference on Data Mining, ICDM. (2008) 995–1000
22. He, J., Gu, H., Liu, W.: Imbalanced multi-modal multi-label learning for subcellular localization prediction of human proteins with both single and multiple sites. PLoS ONE **7** (2012)
23. Li, C., Shi, G.: Improvement of learning algorithm for the multi-instance multi-label RBF neural networks trained with imbalanced samples. Journal of Information Science and Engineering **29** (2013) 765–776
24. Tahir, M.A., Kittler, J., Bouridane, A.: Multilabel classification using heterogeneous ensemble of multi-label classifiers. Pattern Recognition Letters **33** (2012) 513–523
25. Charte, F., Rivera, A.J., del Jesus, M.J., Herrera, F.: Addressing imbalance in multilabel classification: Measures and random resampling algorithms. Neurocomputing **163** (2015) 3–16
26. Giraldo-Forero, A.F., Jaramillo-Garzón, J.A., Ruiz-Muñoz, J.F., Castellanos-Domínguez, C.G.: Managing imbalanced data sets in multi-label problems: A case study with the SMOTE algorithm. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Volume 8258 LNCS. (2013) 334–342
27. Zhang, M.L., Li, Y.K., Liu, X.Y.: Towards class-imbalance aware multi-label learning. In: IJCAI International Joint Conference on Artificial Intelligence. Volume 2015-January. (2015) 4041–4047
28. García, V., Sánchez, J.S., Mollineda, R.A.: On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. Knowledge-Based Systems **25** (2012) 13–21
29. Charte, F., Rivera, A.J., del Jesus, M.J., Herrera, F.: Mlsmote: Approaching imbalanced multilabel learning through synthetic instance generation. Knowledge-Based Systems **89** (2015) 385 – 397
30. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401** (1999) 788–91
31. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics (2010) 100–108
32. Bouma, G.: Normalized ( Pointwise ) Mutual Information in Collocation Extraction. Proceedings of German Society for Computational Linguistics (GSCL 2009) (2009) 31–40
33. Aletras, N., Stevenson, M.: Evaluating topic coherence using distributional semantics. Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers (2013) 13–22
34. O’Callaghan, D., Greene, D., Carthy, J., Cunningham, P.: An Analysis of the Coherence of Descriptors in Topic Modeling. Expert Systems with Applications **13** (2015) 5645–5657
35. Mikolov, T., Corrado, G., Chen, K., Dean, J.: Efficient Estimation of Word Representations in Vector Space. Proceedings of the International Conference on Learning Representations (ICLR 2013) (2013) 1–12
36. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys **34** (2002) 1–47