Distributed multi-Label classification approach for textual Big Data

Houda Amazal, Mohammed Ramdani, Mohamed Kissi

Distributed Multi-Label Classification Approach

For Textual Big Data

Houda Amazal, Mohammed Ramdani, Mohamed Kissi
Houda.kamouss@gmail.com
Computer science laboratory - LIM@II
Faculty of sciences and technologies,
University Hassan II
Mohammedia, Morocco

***Abstract* -- With the increased generation of data, classification still a hot research topic in machine learning. Though a lot of works in literature are interested in single-label classification, the huge amount of dimensionality of data requires new approach. Thus, multi-label classification has attracted significant attention in the research community over the last years. This task which is an extension of the single-label classification, consists of associating an instance of data (document) with multiple labels; which is practical in many domains such as image analysis, bio-informatics, and text categorization, among others. Besides that multi-label classification is a challenging task, the high dimensionality requires the use of distributed environment to manage data effectively and efficiently. Thus, in this work we propose a distributed system to classify documents using Hadoop framework. Documents are given to the MapReduce framework which assigns the set of positive labels to the documents using a distributed approach based on the Label Powerset method. Experiments on real-life data were carried out to show that the proposed approach can effectively reduce redundant attributes and improve multi-label classification accuracy.**

***Index Terms*--** Big Data, Feature Selection, Machine Learning, MapReduce, Multi-label, Text Classification.

I. INTRODUCTION

With the rapid development of Internet technology, huge amounts of textual data are permanently produced, given arise to big textual data phenomenon. Actually, traditional machine learning techniques are unable to classify and consequently to extract useful information hidden within these data. Another challenge to overcome is that a large part of the data has the inability to belong to only one category (label). Definitely, the data can be associated with one or more categories; this data is called multi-label data. The classification of multi-label data is called multi-label classification or multi-label learning [1], [2], [3]. Though multi-label classification is a very difficult classification problem, it has attracted increasing attention in research, due to its high application value; for example, a text document may fall into one of the following categories both in the economy and computer categories, an electronic message may be labelled both as work and research project, and so on.

A lot of supervised learning research are focused on the analysis of single-label data. Single-label classification [4], often known as multi-class classification, is the common machine learning task where a document is assigned a single label l, that is chosen from a previously known finite set of labels L. A dataset D of n documents is composed of document-classification pairs $(d_0, l_0), (d_1, l_1), \dots, (d_n, l_n)$. Multi-label classification [5] is an extension of this problem, where documents are classified with a subset of labels $S \subseteq L$. A multi–label dataset D of n documents is composed of document-classification pairs $(d_0, S_0), (d_1, S_1), \dots, (d_n, S_m)$.

Multi-label classification methods can fall into one of the two main categories: a) problem transformation methods, and b) algorithm adaptation methods [6]. Problem transformation methods consists of transforming the multi-label classification problem either into one or more single-label classification. While algorithm adaptation methods recover methods that extend specific learning algorithms in order to handle multi-label data directly.

In the literature, the majority of the proposed methods address multiple label problems by first transforming a multiple label problem into a set of independent binary classification problems, and then using thresholding for multi-label classification. One of the most used technique is the Binary Relevance Method (BR) [7], which treats each class as a separate binary classification problem. The BR method has several advantages. It is theoretically simple, intuitive and generally not very complex from a computer point of view, but this approach ignores the dependencies between labels. For this, another popular transformation method called Label Powerset (LP) [8] is preferred. This method creates its single-label problem simply by treating each document's label subset S_i as a single label l. The set of all distinct labels is used as the set of possible labels L for a single-label classifier to choose from. For example, a document (d, S) where S = {a, c, d} is transformed into the single-label representation (d, l) where l = acd. The reverse transformation is evident. The total number of classes the single-label classifier must learn is the total number of distinct label sets found in the training set. But, as the possible combinations of categories are 2^L , where L is the number of separate labels in the data set, this method may lead to data sets with a large number of classes, the resulting power set data tends to become scarce and

therefore makes the classifier's work more difficult [9]. To deal with this, the original Label Powerset technique has been extended and improved. One of its variants is Pruned Problem Transformation (PPT), proposed by [10], which eliminates label sets that occur below a certain minimum threshold value, in our work we adopted this method as Pruned Label Powerset 1 (PLP1). However, clearly a lot of information is potentially lost in this approach. For this another approach we noted PLP2 is proposed. It consists of splitting the combination subset S_i, into k sub-subsets $(S_{i0}, S_{i0}, .., S_{ik})$ where these sub-subsets do occur more than x times in the training data.

In this work we have chosen the TFIDF technique [11], which is one of the most well-known method for feature selection, and we have evaluated it along with a common transformation technique for the multi-label classification. We have also adjusted the TFIDF feature selection technique to handle large multi-label data set directly using Hadoop cluster. To evaluate the performance of this work, parallel Naïve Bayes [12] classifier is used. Our goal is to perform a comprehensive study of the performance of distributed multi-label feature selection techniques based on document frequency and report how it varies when coupled with two different multi-label approaches using large data sets.

The remainder of this paper is organized as follows. In Section II, we introduce related works. In Section III, we describe the multi-label classification process we adopted. In Section IV, we report the experiments and analysis. Finally, in Section V, we make our concluding remarks and point to directions for future research.

II. RELATED WORKS

Multi-label classification is in very high demand because of its fast growing requirement in applications. In [13] authors evaluate the performance of information gain for multi-label classification problem. The work adapted the common transformation methods then employ feature selection measure to select the best features. The goal being to report the performance of multi-label feature selection techniques using information gain measure and state variations when coupled with different multi-label classifiers and data sets.

In [14], an algorithm based on MapReduce is used to assign labels to web documents. The proposed algorithm assigns the set of positive labels to the documents of the web using binary classification of binary classifier. The proposed approach satisfactorily classifies the labels to the documents of the web.

In [15], authors introduced a new fine-grained weight method to improve the performance of the multinomial naive Bayes classification of documents. The idea of fine-grained weight method is to assign a different weight to each world frequency. The result show that the new weighting method could significantly improve the performance of multinomial nave Bayes learning.

In [16], an extension of the previous work is presented. Thus, to improve multi-label classification of documents using fine-grained weight method, this work uses the dependencies between labels.

In [17], an algorithm for Multi-label learning with label specific features is introduced. The algorithm does clustering analysis on positive and negative patterns and based on that, constructs label specific features. Then it queries the clustering results to perform training and testing.

Multi-label Classification Using Ensembles of Pruned Sets is introduced in [18], this pruned set method (PS) treats sets of labels as single labels. First it converts the multi-label problem into a single label problem. This way the correlation between labels are taken into account. After this, corresponding to each training instance, it considers the label sets as a single class if the label set occurs more than a threshold value. Otherwise the label set is broken into smaller subsets so that it occurs more than the threshold.

In [19], authors attempt to improve the performance and accuracy of a multi-label classification algorithm for large data sets, using parallel computing in a distributed manner. The proposed algorithm can reduce the dimensionality of large data sets with very large number of features by removing the redundant features using a feature selection method (Fscore) [20] to improve the accuracy and reduce the time taken for training phase of the multi-label classification algorithm.

It can be noted that a lot of works in literature are developed using no distributed environment, this constitutes a limitation of these works since they cannot handle all data for experiments and this will affect their results. The contribution of our research can be resumed on elaborating a parallel system for multi-label classification using parallel Naïve Bayes based on document-based feature selection technique coupled to two variations of Pruned Problem Transformation.

III. METHODOLOGY

This section presents our complete methodology to perform feature selection for multi-label classification problems. Our data set is divided into training data and testing data. Then training data are preprocessed and each document is assigned to its categories. In the PLP1 method, documents belonging to categories with less than a threshold value are eliminated. While in the PLP2 method this kind of documents are assigned to a sub-categories. Then documents are loaded in HDFS. Next, the TFIDF score of each feature is calculated using MapReduce, and features of each document are ranked according to their scores. Features having score superior to a threshold value are indexed as relevant and all others are eliminated. The most relevant features are used by a parallel implementation of Naïve Bayes to build the classifier. Then, test documents are introduced to be classified using the previous classifier. These steps are described in Fig.1.

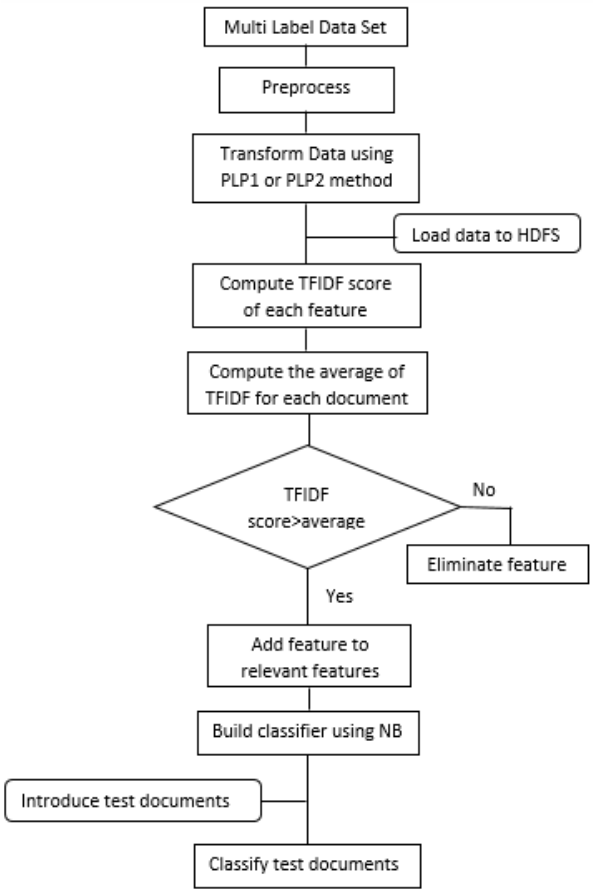

Fig. 1. Flowchart of the proposed approach

IV. EXPERIMENTS

In this paper, we use the OHSUMED dataset [21] for evaluating the performance of our parallel document-based filter feature selection approach. The dataset used is composed of medical abstracts from 270 medical journals over a five-year period (1987-1991) categorized into 23 cardiovascular diseases categories. The detailed description of this dataset is presented in Table 1.

Table 1
The description of Ohsumed dataset

Category	Train doc	Test doc
C01	423	506
C04	1163	1467
C06	588	632
C08	473	600
C10	621	941
C12	491	548
C14	1249	1301
C20	525	695
C21	546	717
C23	1799	777

At first, the data set includes 181 categories, this huge number of categories was reduced to 54 categories using the PLP1 method. Categories used are obtained by removing categories represented by less than a threshold which is 10 documents, given that the category represented by the largest number of documents include 723 instances. This steps leads to delete 319 documents from training phase which will certainly affects classification performance since a lot of information will be lost. To achieve better performance further experiments are needed to determine the optimal threshold to use for retrieving categories. As another alternative approach, the PLP2 method is used. This method is an upgrading of the PLP1 method where all documents of the training phase are used, thus no training document will be removed. The results of classification using PLP1 and PLP2 methods are given in Table 2.

Table 2
The performance of our approach using the PLP1 and PLP2 methods

Category	**Mesaure**	**PLP1**	**PLP2**
C01	Precision	27.08%	31.82%
	Recall	28.19%	32.66%
	F1-measure	27.62%	32.23%
C04	Precision	39.95%	42.47%
	Recall	39.14%	41.42%
	F1-measure	39.54%	41.94%
C06	Precision	22.63%	27.53%
	Recall	23.37%	28.48%
	F1-measure	22.99%	28.00%
C08	Precision	29.67%	35.00%
	Recall	30.17%	34.54%
	F1-measure	29.92%	34.77%
C10	Precision	29.86%	32.73%
	Recall	29.18%	31.92%
	F1-measure	29.52%	32.32%
C12	Precision	29.01%	33.39%
	Recall	27.99%	32.62%
	F1-measure	28.49%	33.00%
C14	Precision	45.50%	50.12%
	Recall	45.68%	53.66%
	F1-measure	45.59%	51.83%
C20	Precision	29.78%	34.10%
	Recall	30.53%	34.35%
	F1-measure	30.15%	34.22%
C21	Precision	32.78%	36.40%
	Recall	32.32%	34.94%
	F1-measure	32.55%	35.66%
C23	Precision	28.06%	31.66%
	Recall	28.42%	31.14%
	F1-measure	28.24%	31.40%

The experimental steps are as following:

- For preprocessing, stop words are removed. The stemming process is executed with Porter Stemming algorithm (Porter 1997).

- For text representation we use bag-of-words technique.
- The PLP1 and PLP2 methods are used separately.
- The training bag-of-words vectors are reduced by a parallel TF-IDF.
- Then, they are used for building a leaning model using parallel NB classifier

All experiments were carried out on a 2.70GHz Intel(R) Core i7-4600U Processor with 4 gigabytes of memory running Ubuntu 16.4. This environment used in a multi nodes Hadoop cluster including 5 nodes.

It can be noted from Table 2 that both PLP1 and PLP2 have a very low accuracy rate 31,43% and 35,52%Respectively. This can be due to the document based feature selection technique which ignores relation between features and labels and also to the imbalanced aspect of data. Effectively this can explains the difference of precision among some categories; for instance C04 and C14 have the highest scores and the highest number of documents in the data set.

Fig.2. Shows that PLP2 method perform better than PLP1 for all measures, this can be due to the fact that this approach uses all training documents.

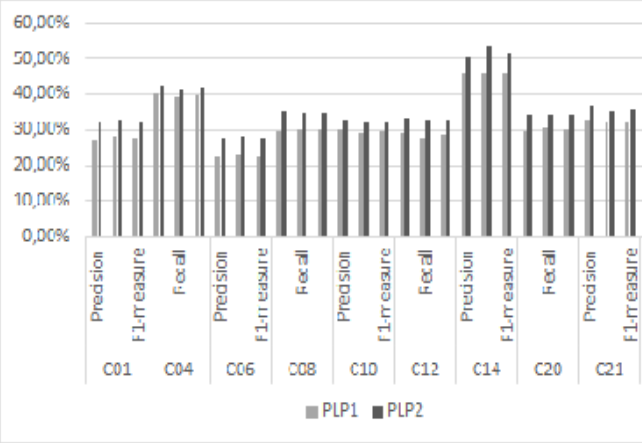

Fig. 2. Comparison of measures for PLP1 and PLP2 methods

V. CONCLUSION

In this paper a parallel feature selection technique coupled to a pruned problem transformation method to incorporate the dependencies among the class values is used to improve multi-label classification and reduce space dimensionality. A parallel Naïve Bayes classifier is used to evaluate the performance of the selected features. Parallel algorithms are implemented using different number of nodes in a Hadoop cluster. The results of the experiment show that both feature selection techniques and transformation methods affect considerably the performance of multi-label classification. For the future, it will also be interesting to use class-based feature selection techniques to better benefit from the relationships between labels and introducing other measures related to multi-label classification. Using other large data sets for test will also be suitable.

VI. REFERENCES

[1] Gibaja E, Ventura S (2015) A tutorial on multilabel learning. ACM Comput Surv 47(3):1–38

[2] Tsoumakas G, Katakis I (2007) Multi-label classification: an overview. Int J Data Warehouse Min 3(3):1–13

[3] Streich AP, Buhmann JM (2008) Classification of multi-labeled data: a generative approach. Mach Learn Knowl Discov Databases DBLP:390–405

[4] Li, Y., Li, T., & Liu, H. (2017). Recent advances in feature selection and its applications. *Knowledge and Information Systems*, *53*(3), 551-577.

[5] Pereira, R. B., Plastino, A., Zadrozny, B., & Merschmann, L. H. (2018). Categorizing feature selection methods for multi-label classification. *Artificial Intelligence Review*, *49*(1), 57-78.

[6] Pant, P., Sabitha, A. S., Choudhury, T., & Dhingra, P. (2019). Multi-label Classification Trending Challenges and Approaches. In *Emerging Trends in Expert Applications and Security* (pp. 433-444). Springer, Singapore.

[7] Kashef, S., Nezamabadi-pour, H., & Nikpour, B. (2018). Multilabel feature selection: A comprehensive review and guiding experiments. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(2), e1240.

[8] Abdallah, Z., El-Zaart, A., & Oueidat, M. (2016). An Improvement of Label PowerSet Method Based on Priority Label Transformation. *International Journal of Applied Engineering Research*, *11*(16), 9079-9087.

[9] Dembczynski, K., Kotlowski, W., & Hüllermeier, E. (2012). Consistent multilabel ranking through univariate losses. *arXiv preprint arXiv:1206.6401*.

[10] Read, J. (2008, April). A pruned problem transformation method for multi-label classification. In *Proc. 2008 New Zealand Computer Science Research Student Conference (NZCSRS 2008)* (Vol. 143150).

[11] CHASE, Z., Genain, N., & Karniol-Tambour, O. (2014). Learning Multi-Label Topic Classification of News Articles.

[12] Amazal, H., Ramdani, M., & Kissi, M. (2018, October). A Text Classification Approach using Parallel Naive Bayes in Big Data Context. In *Proceedings of the 12th International Conference on Intelligent Systems: Theories and Applications* (p. 36). ACM.

[13] Pereira, R. B., Plastino, A., Zadrozny, B., & Merschmann, L. H. (2015). Information gain feature selection for multi-label classification. *Journal of Information and Data Management*, *6*(1), 48.

[14] Malarvizhi, P., & Pujeri, R. V. (2013). Multilabel classification of documents with mapreduce. *International Journal of Engineering and Technology*, *5*, 1260-1267.

[15] S.-H. Song and C.-H. Lee, Improving Multi-label Classification of Documents Using Fine-Grained Weights. in IEA/AIE 2015: The 28th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, 2015

[16] Lee, C. H. (2018). Multi-label classification of documents using fine-grained weights and modified co-training. *Intelligent Data Analysis*, *22*(1), 103-115.

[17] M.-L. Zhang. Lift: multi-label learning with label-specific features. in Proceedings of the Twenty-Second international joint conference on Artificial Intelligence, AAAI Press, Volume Two, ser. IJCAI11.:1609–1614, 2011.

[18] J. Read, B. Pfahringer, and G. Holmes. Multi-label classification using ensembles of pruned sets. in Data Mining, 2008. ICDM 08. Eighth IEEE International Conference on, pages 995 –1000, Dec. 2008.

[19] Biswas, S., & Devi, V. S. (2018, November). Parallelization of Multi-label classification for large data sets. In 2018 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 2005-2010). IEEE.

[20] Murty, M. N., & Devi, V. S. (2015). *Introduction to pattern recognition and machine learning*.

[21] Le, N. H. N., & Ho, B. Q. (2015). A comprehensive filter feature selection for improving document classification. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation* (pp. 169-177).

[22] Alboaneen, D. A., Tianfield, H., & Zhang, Y. (2017, December). Sentiment analysis via multi-layer perceptron trained by meta-heuristic optimisation. In 2017 IEEE International Conference on Big Data (Big Data) (pp. 4630-4635). IEEE.