DV 2542 - MACHINE LEARNING PROJECT

COMBINING CLUSTERING AND MULTI LABEL TEXT CLASSIFICATION FOR TAG SUGGESTION

K S SUSHEEL SAGAR
9303177837
suko15@student.bth.se

I. INTRODUCTION

Text classification is a process, that involves in assigning one or more predefined categories to a free text according to its content[6]. Filtering of Spam e-mails, Information Retrieval from large Databases, user's preference based searching of data are some examples where text categorization play a vital role.

Traditional single-label classification is concerned with learning from a set of data, where each instance in the data is associated with a single label. In several text classification applications however, each instance in the data is associated with a set of label[3]. Tagging can be defined as the process of assigning short textual descriptions or keywords (called tags) to information objects. Within most of the web 2.0 applications tag suggestion play an important role to quickly assign the tags to the data[4].

GOAL : To increase the classification accuracy of the multi label text classification process by using K-means clustering.

MOTIVATION : Recent studies on combining clustering and classification proved to attain improved accuracy for the domain of software quality evaluation[1]. This motivated me to check the impact of combining the clustering and multi label text classification tasks for the tag selection process.

Project Plan : The project plan presented in the table below, explains the approximated time, spent on each task of this project.

TASK	Number of Hours Spent
Initial Idea about Task	30
Literature Review based on Initial Idea	40
Re Analysing the Initial Idea	15
Searching for the Relevant Software/Tools	5
Searching and obtaining the Dataset	12
Initial configuration of Apis	5
Coding the Idea	20
Analyzing the Implementation	5-8
Documenting the Project	20
TOTAL	155(approximately)

II. LITERATURE REVIEW

A literature study was done in-order to obtain knowledge about the multi label learning algorithms and its applications for text classification. Various Databases such as Inspec, IEEE explore and google scholar were searched for the articles related to Multi Label Learning and the combination of clustering and classification. The table presented below explains about the articles that are important for this project.

REFERENCE	CONTENT
[1]	This article explains that the combination of clustering and classification provides better accuracy that just clustering
[2]	This articles explains about the software package that has implementations of Multi label learning algorithms
[4]	This articles explains about the modeling of automated tag selection problem as a multi label classification problem.

[5]	This article compares several multi label learning algorithms and presents performance of each of the algorithms on different datasets.
[8]	This article proposes a multi label learning algorithm called MLkNN and states that, this algorithm outperforms several other state of the art algorithms.

III. IMPLEMENTATION

A. DataSet :

For this project a multi label classification data set that has text data is required. For this purpose, “bibtex” dataset from the MULAN repository was selected. The table presented below explains about the dataset [7].

PROPERTY	VALUE
Number Of Instances	7395
Domain	TEXT
Nominal Attributes	1836
Numeric Attributes	0
Labels	159
Cardinality	2.402

B. Algorithms Selection:

In order to compare a Multi label classification algorithm with a combination of a clustering algorithm and a Multi label algorithm, we need to select 1 multi label learning algorithm and 1 clustering algorithm. For our study we chose MLkNN [8] for multi label classification and k-means Algorithm for Clustering.

Motivation behind selecting MLkNN: The prime motivation behind choosing MLkNN is that it has a low training time and is proved to be outperforming several well established multi label learning algorithms [8].

C. Mulan and Weka:

Mulan is an open source java library for learning from multi label datasets. The Library includes a variety of state of the art algorithms for performing the major multi label learning tasks such as Classification, Ranking, Classification and Ranking. In addition, The Library also offers the functionalities such as feature selection and Evaluation[7].

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from a java code. Weka contains tools for data pre-processing, classification,regression, clustering, association rules and visualization.

D. Coding :

The data set is initially available in the arff format, and hence no preprocessing is required and is readily available for the learning algorithms. The MLkNN algorithm from the Mulan package is directly used for the multi label classification. While the k-means clustering algorithm from weka package is directly used for clustering.

E. Evaluation Metric :

Accuracy is one of the widely considered metric for evaluating classification algorithms. However, for the evaluation of the multi label learning algorithms, several evaluation metrics such as Classification accuracy or subset accuracy , Precision, Recall, F-Measure etc., are used as the measures by several researchers in [9][10][18]. For our study we chose classification accuracy as the evaluation metric.

Motivation for choosing Classification Accuracy:

The motivation behind choosing classification accuracy as the evaluation metric is that, it is a very strict evaluation metric as it requires the ***predicted set of labels to be an exact match of the true set of labels***[9][10].

IV EXPERIMENTATION:

Design Choices :

1. The value of k in the k- means algorithm is set to 4 based on the random choice, with an intention to increase based on the results obtained with k=4.
2. The evaluation of the algorithm is based on the 10-fold cross validation.

To realize the project goal, two experiments must be performed.

A) **Experiment 1** :- In the first experiment, the bibtex data(bibtex.arff) is directly classified using the MLkNN algorithm in Mulan. The evaluation of the model is

done using the 10 fold cross validator in the mulan package and the results are noted down.

B) **Experiment 2** :- In the second experiment, the bibtex data is first clustered using the k-means algorithm (with k= 4). Then the MLkNN classifier from Mulan package is applied on each of the clusters, and their respective 10-fold cross validation results are noted down.

V Results and Analysis:

Experiment	**Instances**	**Classification Accuracy**
MLkNN on entire Dataset (Before Clustering)	7395	0.0617

MLkNN on cluster0	1040 (n_1)	0.0173(C_1)
MLkNN on cluster1	1341(n_2)	0.0664(C_2)
MLkNN on cluster2	3650(n_3)	0.0929(C_3)
MLkNN on cluster3	1364(n_4)	0.0851(C_4)
Weighted Average of Classification Accuracy		0.0760

Weighted Average of Classification Accuracy = $(n_1*C_1)+(n_2*C_2)+(n_3*C_3)+(n_4*C_4) / (n_1+n_2+n_3+n_4)$

The results clearly show that the Classification Accuracy increased upon initial clustering of the multi label text data.

VI LIMITATIONS:

The following are the limitations of the project:

- Only one clustering and multi label classification algorithms are used for the experiment.
- The experiments are run on a single dataset and the results might differ for different datasets.
- The number of clusters is chosen as 4 (randomly). However an exhaustive experimentation must be conducted with the value of k while determining the number of clusters.

VII Conclusion :

Based on the results obtained from the experiments, we can infer that the classification accuracy of the multi label classification algorithm can be improved with the initial clustering of the datasets. The reason for this is that the initial clustering divides the data into more homogeneous groups and when the classification algorithm is applied on these homogeneous groups the classification accuracy is increased.

However, previous studies on combining clustering and classification have shown the increase in accuracy to a significantly large extent [1]. But in the present context as we are not dealing with the accuracy, instead the classification accuracy which is considered in several multi label text classification algorithms(explained in section III(E)) must be studied in greater detail, in order to identify the percentage increase of this measure with the combination of clustering and classification.

VIII References :

[1] Papas, Diomidis, and Christos Tjortjis. "Combining Clustering and Classification for Software Quality Evaluation." *Artificial Intelligence: Methods and Applications*. Springer International Publishing, 2014. 273-286.

[2] Tsoumakas Grigorios, et al. "Mulan: A java library for multi-label learning."*The Journal of Machine Learning Research* 12 (2011): 2411-2414.

[3] Tsoumakas, Grigorios, Ioannis Katakis, and Ioannis Vlahavas. "Random k-labelsets for multilabel classification." *Knowledge and Data Engineering, IEEE Transactions on* 23.7 (2011): 1079-1089.

[4] Tsoumakas Grigorios, Ioannis Katakis, and Ioannis Vlahavas. “Multi Label Text Classification for automated tag suggestion." *International Workshop at the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases in Antwerp,* Belgium, 2008.

[5] Zhang, Min-Ling, and Zhi-Hua Zhou. "A review on multi-label learning algorithms." *Knowledge and Data Engineering, IEEE Transactions on* 26.8 (2014): 1819-1837.

[6] Qin, Yu-ping, and Xiu-kun Wang. "Study on multi-label text classification based on SVM." *Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on*. Vol. 1. IEEE, 2009.

[7] http://mulan.sourceforge.net/datasets-mlc.html

[8] Zhang, Min-Ling, and Zhi-Hua Zhou. "ML-KNN: A lazy learning approach to multi-label learning." *Pattern recognition* 40.7 (2007): 2038-2048.

[9] Zhu, S., Ji, X., Xu, W., Gong, Y.: Multi-labelled classification using maximum entropy method. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in Information Retrieval. (2005) 274–281

[10] Ghamrawi, N., McCallum, A.: Collective multi-label classification. In: Proceedings of the 2005 ACM Conference on Information and Knowledge Management (CIKM '05), Bremen,Germany (2005) 195–200

[11] Godbole, S., Sarawagi, S.: Discriminative methods for multi-labeled classification. In: Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2004). (2004) 22–30