Data Science and Reinforcement Learning (AAI5002.01-00)

Final Exam

Wonpyo Hong 홍원표
2021311625
2021-06-17

Document Categorization via Matching and Clustering

In Today world, numerous digital documents are achieved by each industry every day. As these data can be keystone of further development and evaluation of the industry, data analysis has been one of the primary investments in businesses. For an accurate data analysis, document categorization has been a fundamental task to be solved.

Document Categorization task had been approached by several methods using classification, which attempt to solve document categorization using partially labeled documents for training classifiers, known as semi-supervised text classification. Based on the idea that small amount of trainable data can lead to generalization of unlabeled data, these approaches were able to show prominent performances thanks to the advancement in text classification methods like BERT.

However, these approaches have several challenges to be applied to real industry data, which can be generalized as follows:

- **[C1] Cold-start Categorization**: Excluding few technologically developed industries where pre-labeled data are present, industrial data do not provide seed information to guide proper categorization. Also, acquiring even small amount of labeled data from domain experts is expensive, considering the enormous amount of data to deal with.
- **[C2] Inaccurate pseudo categories**: Pseudo categories may not always be correct; pseudo categories based on hasty generalization of entire dataset may lead to misguidance in final output. Pseudo categories refer to the keywords that can be generated by skimming the data through, not matched to specific documents. For example, if the given documents are news articles, one could easily tell that some of the categories would be *businesses, politics, or sports*. Hence, it is inevitable that pseudo labels do not sufficiently reflect the entire dataset.

Naïve methods to tackle **[C1]** such as weakly-supervised classifications have been proposed. These approaches, however, have limitations in solving **[C2]** as they assume pseudo categories to be directly related to output labels, as it has un-scalable representation dimensions, thus resulting

a biased result that can easily sway depending on the decision of initial pseudo categories. Weakly-supervised methods of document classifications can certainly be useful for industrial datasets with gold labels to compare and compute loss with, or with partially labeled documents that can act as a ground indicator to train with. However, this also implies that a method relying solely on classification is like a fool's gold, since it is not capable of categorizing datasets in real world, where a significant amount of them are cold-start datasets.

In contrast, in this paper, we propose a more systematic approach of integrating both **[C1]** and **[C2]** into one, by solving them independently and integrating them together as a pipeline. Our proposed method seeks solutions to use pseudo labels only up to training classifiers, and then implements document clustering to form a new set of labels as a result. Clustering diminishes the impact of seed information of pseudo labels upon the result categories. Our method, named CMC, for Categorization via Matching and Clustering, contains 4 separate procedures: (1) Pseudo Label Matching, (2) Classifier Training, (3) Clustering, and (4) Keyword Extracting.

Our proposed method (1) selects a portion of documents that are most relevant to pseudo labels by matching semantic relationships between the labels and documents. These matched documents behave as the partially labeled documents for training classifiers as the process of semi-supervised classification methods. After (2) training classifiers, we retrieve the final representation of classifier test predictions. From this point, the method takes the representation (3) to construct graph structures which facilitates formation of clusters. These clusters indicate newly formed categories with unknown category name; thus we (4) perform keyword extracting methods per each cluster to obtain the most representative keyword.

For evaluation and comparisons with baseline methods, we use public datasets such as DBPEDIA and 20Newsgroup. Since pseudo categories do not exist in these datasets, we perform some alteration to original set of given labels and obtain a set that would act as a similar role of pseudo categories.

In summary, the contribution of the paper are listed below:

- We propose a categorization method which do not require any labeled documents, thus minimizing human efforts. Our proposed method, named CMC for Categorizing via Matching and Clustering, combines only the advantages of document classification and clustering to facilitate cold-start categorization where gold-labels are missing.
- We propose an architecture that can be separated into four different procedures, where each procedure is model agnostic. We introduce some of the popular models that can be implemented in each procedure, then suggest the most befitting model.
- We propose a new method to analyze the accuracy of clustering result, considering that there has not been an evaluation method for cold-start datasets to the best of our knowledge.