University of Essex

CE807 – Assignment 2 - TASK 1: XMLC literature Review

Registration Number: 1901423

University of Essex – Colchester Campus
Department: Computer Science and Electronic Engineering
CE807 – Assignment 2 - Final Practical Text Analytics and Report
Dr Habil Ansgar Scherp(Supervisor)

April 21, 2020

1 Introduction

The text classification problem has been extensively discussed in data analysis, natural language processing, database, and machine vision with implementations being seen in various fields such as target advertising, medical diagnosis, message filtering, and data organization. Text classification (TC) or simply text categorization is a mechanism whereby one or more predefined groups are automatically assigned to a text document.[1]

Almost all the TC research activities have been focusing on the flat classification in which the predefined categories or subgroups are treated in seclusion and no structure defines relationships between them. However, as the number of groups increases to a considerably large level, it becomes much harder to scan and check the groups.

The primary focus is on discussing the varieties of algorithms and methods for classifying text. Most of the recent projects on classification is based on strategies that only involve a set of explicitly categorized training cases that are far less costly to produce. A classifier is constructed by training from a set of pre-classified instances.[2]

One downside to supervised text classification methods is that they must be trained on specified positive and negative test cases or predefined classes. The accuracy of these classification algorithms depends on the quality of the sample tests used. With the vast volumes of data and various types of applications, creating the training sets or the contextual classes is not easy to achieve manually.[3]

2 Classification Methods

In general, two main methodologies are utilized in text classification: rule-based and machine learning-based methodologies. The rule-based method is a text classification technique in which the categorization rules are explicitly (or manually) defined and the data elements are categorized based on rules.

On the other hand, Machine learning methods refer to the text classification techniques where the categorization rules or matrices utilizing sample labeled data. The Machine learning-based methods are more efficient but have slightly lower precision than the rule-based methods. Due to their efficiency in terms of memory usage and execution cost, the Machine-based learning methods are preferred and are fast-replacing the rule-based methods in text classification.[4]

The machine learning based techniques are further divided into two broad categories – the unsupervised and supervised learning. [5] Supervised learning method is where a predefined class labels are provided for the training data set, while, an unsupervised text classification (also known as text clustering) is where the categorization (or classification) is done implicitly without referring to any to external information (or a predefined knowledge base).

To limit the scope of this study, only one supervised learning model will be considered, the extreme multi-label classification method which is yet another generalization of the conventional Multi-class-based learning models. When considering the multi-label case, the classes are not inherently exclusive, and any test set can belong to multiple classes simultaneously.[6]

3 Extreme Multi-Label Text Classification (XMC)

The Extreme multi-label text classification (XMC) is a categorization scheme that attempts to tag each text with the most appropriate subset of classes (labels) from a large collection of labels, where the size of the dataset might be in the millions. [7] In simple terms, the XMC method aims at tagging input training data (which is a text) with the most appropriate labels from an enormous data set.

In recent times, XMC has attracted a worthwhile attention because of the rapid growth of data due to extensive internet usage (also called web-scale data) that is present in different systems, such as e-commerce product categorization, Bing's dynamic query adverts [4, 10], and labeling of Wikipedia sections and categories in the PASCAL Large-Scale Hierarchical Text Classification (LSHTC) problem, just to outline a few. XMC algorithms can be broken down into several distinct categories: deep learning-based methods, partitioning methods, and One-vs-All methods.[8]

3.1 One-vs-All (OVA)Approaches

The One-vs-All (OVA) strategy considers every label individually as a binary categorization problem. It has been proven that the OVA methods do provide high precision, however, when the dataset (labels) becomes very large, they strain from costly computation for both model training and prediction.[9]

Consequently, several strategies have been introduced to optimize the algorithm to ensure a low computation cost. The two popular types of OVA are Primal and Dual Sparse (PD-Sparse), Distributed Sparse Machines for Extreme Multi-label Classification (DiSMEC) algorithms.[[4], [8],[10], [11]]

PD-Sparse uses the concept of primal and dual computational complexity (or sparsity) to improve both the model training and prediction. Contrary to the mechanisms employed other state-of-the-art approaches, DiSMEC makes no low-rank assumptions the matrix classes. DiSMEC is capable of learning classifiers for datasets consisting of millions of labels within a short time using a double layer parallelization mechanism.

The supervised capacity control technique filters out arbitrary parameters thereby keeping the model compact in size, to maintain is predictive accuracy. Besides, in order to improve the efficiency of the algorithm and reduce the training set size, DiSMEC explores parallelism and sparsity.

Moreover, the OVA techniques are commonly utilized as building blocks for many other text classification methods. Even though OVA schemes are always considered efficient they do suffer from skewed distributions (i.e. class imbalance). The classifiers used in making predictions struggle to handle imbalance datasets. An imbalance dataset are those with samples of varying sizes (i.e. Class Y makes 84% of the data and samples of another class say Z makes only 16%) [12].

3.2 Partitioning Methods

The most widely used partition strategy for clustering in text classification is k-means due to its effectiveness and efficiency in achieving predictions even when analyzing enormous sparse matrices. When considering the two-way dataset (object x variable) K-means algorithm utilizes the one-way clustering strategy with the ultimate goal being the classification of the objects. [9]

Besides, it has been established that the best approach to use when considering a cost-effective k-means algorithm is co-clustering which entails simultaneous partitioning of rows and columns; the key concept is to label the sub-matrices of the training set, in which each block defines a cluster of objects and variables. The significance of this method is its ability to find clusters of the text identified by categories of terms.

Because of its vast important features, the Co-clustering technique is used in a variety of studies that involve multiple attribute analysis. It is worth pointing out that the study of co-clustering was introduced in text mining to address the issue of multi-partition of texts in the digital library since it is very effective in detecting the repetition of terms and texts in the same dataset. The main drawback of the Partitioning based methods is that are inefficient when working with the large datasets because of the amount of memory they consume when processing large volumes of datasets.[13]

3.3 Deep Learning-Based Methods

Deep learning employs a collection of models, algorithms, and techniques to simulate the workings of the human brain [5]. Thanks to their ability to achieve high precision with the use of slightly less engineered features, the field of text classification have significantly thrived from the resurgence of deep-learning algorithms.

Two main deep learning algorithms are utilized in text classification – the Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). CNN is a very efficient and powerful artificial neuro network algorithm.It preserves the spatial structure of the problem and were developed for object recognition tasks such as handwritten digit recognition. [14] Recurrent Neural Network is an artificial neural network and it involves connection of nodes from a given directed graph along a given pattern allowing for the exhibition of dynamic temporal behavior for a given sequence.

One of the drawbacks that are associated with these algorithms is their need for enormous training data. For example, the algorithms require at least millions of labeled instances. Even though deep learning algorithms require enormous amount of training sets to be efficient, they are still considered to be the best option when performing text classification.

4 Conclusion

Handling and processing unstructured data continue to create several challenges in almost all data-intensive application areas such as business setups, learning institutions, and technology-intensive firms.

It is said that approximately 80% (eighty percent) of an entity's data (such as a person) are available in an unorganized form. Text mining (also known as text analytics) analyzes the secret relationships between entities in the data sets so that meaningful patterns that represent the information found in the data sets can be inferred. The inferred information or knowledge can then be used in informed decision making. In this article, some of the most popular XMC text classification algorithms are discussed.

References

[1] M. Yan, "Adaptive learning knowledge networks for few-shot learning," *IEEE Access*, vol. 7, pp. 119041–119051, 2019.

[2] L. Galke, F. Mai, A. Schelten, D. Brunsch, and A. Scherp, "Using titles vs. full-text as source for automated semantic document annotation," in *Proceedings of the Knowledge Capture Conference*, pp. 1–4, 2017.

[3] F. M. Ba-Alwi and M. Albared, "Experiments on the use of machine learning classification methods in online crime text filtering and classification," *Current Journal of Applied Science and Technology*, pp. 1–12, 2016.

[4] J. Gong, Z. Teng, Q. Teng, H. Zhang, L. Du, S. Chen, M. Z. A. Bhuiyan, J. Li, M. Liu, and H. Ma, "Hierarchical graph transformer-based deep learning model for large-scale multi-label text classification," *IEEE Access*, vol. 8, pp. 30885–30896, 2020.

[5] S. Chowdhury and S. Gupta, "A comparative study of text mining in big data analytics using deep learning and other machine learning algorithms," *International Journal of Hybrid Intelligence*, vol. 1, no. 2-3, pp. 163–177.

[6] A. Law and A. Ghosh, "Multi-label classification using a cascade of stacked autoencoder and extreme learning machines," *Neurocomputing*, vol. 358, pp. 222–234, 2019.

[7] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, 2019.

[8] F. Fanny, Y. Muliono, and F. Tanzil, "A comparison of text classification methods k-nn, naïve bayes, and support vector machine for news classification," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 3, no. 2, pp. 157–160, 2018.

[9] P. Prajapati and A. Thakkar, "Extreme multi-label learning: A large scale classification approach in machine learning," *Journal of Information and Optimization Sciences*, vol. 40, no. 4, pp. 983–1001, 2019.

[10] F. Gargiulo, S. Silvestri, M. Ciampi, and G. De Pietro, "Deep neural network for hierarchical extreme multi-label text classification," *Applied Soft Computing*, vol. 79, pp. 125–138, 2019.

[11] M. Motwani and A. Tiwari, "Comparative study and analysis of supervised and unsupervised term weighting methods on text classification," *International Journal of Computer Applications*, vol. 68, no. 10, 2013.

[12] M. Boroš and J. Maršík, "Multi-label text classification via ensemble techniques," *International Journal of Computer and Communication Engineering*, vol. 1, no. 1, pp. 62–65, 2012.

[13] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, "Comparing automated text classification methods," *International Journal of Research in Marketing*, vol. 36, no. 1, pp. 20–38, 2019.

[14] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 7370–7377, 2019.