Semi Supervised Learning Based Text Classification Model for Multi Label Paradigm

Shweta C. Dharmadhikari¹, Maya Ingle², and Parag Kulkarni³

¹Pune Institute of Computer Technology, Pune, Maharashtra, India
²Devi Ahilya Vishwa Vidyalaya, Indore, Madhya Pradesh, India
³EkLat Solutions Pune, Maharashtra, India
{d.shweta18,paragindia}@gmail.com, maya_ingle@rediffmail.com

Abstract. Automatic text categorization (ATC) is a prominent research area within Information retrieval. Through this paper a classification model for ATC in multi-label domain is discussed. We are proposing a new multi label text classification model for assigning more relevant set of categories to every input text document. Our model is greatly influenced by graph based framework and Semi supervised learning. We demonstrate the effectiveness of our model using Enron, Slashdot, BibTex and RCV1 datasets. We also compare performance of our model with few popular existing supervised techniques. Our experimental results indicate that the use of Semi Supervised Learning in multi label text classification greatly improves the decision making capability of classifier.

Keywords: Automatic text categorization, Multi-label text classification, graph based framework, semi supervised learning.

1 Introduction

Automatic text classification (ATC) is a prominent research area within Information retrieval. Multi label text classification problem refers to the scenario in which a text document can be assigned to more than one classes simultaneously during the process of text classification. The inherent ambiguity present in the content of textual data often makes the text document to be the member of more than one class simultaneously[3]. It has attracted significant attention from lot of researchers for playing crucial role in many applications such as web page classification, classification of news articles, information retrieval etc. Multi label text classifier can be realized by using supervised, unsupervised and semi supervised methods of machine learning. In supervised methods only labeled text data is needed for training. But availability of labeled data all the time is rare and processing of is expensive. Unsupervised methods relies only on unlabeled text documents; but it does not shows remarkable improvement in the performance. Semi supervised methods effectively uses unlabeled data in addition to the labeled data. Majority of existing approaches are supervised in nature[16]. Most of these lacking in considering relationship...
between class labels, input documents and also relying on labeled data all the time for classification. And also not capable of utilizing information conveyed by unlabeled data[17].

Hence through our paper we are proposing a multi label classification model using semi supervised learning so that classifier can handle labeled and unlabeled data. We are also aiming at handling input documents similarity along with correlation existing between class labels to improve decision making capability of our proposed classifier. We apply the proposed model on standard dataset such as Enron, Bibtex and RCV1 and Slashdot to test the performance. We also compare performance of our model with few popular existing supervised techniques.

The rest of the paper is organized as below. Section 2 describes relevant literature related to our proposed system; Section 3 describes our proposed classification model. Section 4 describes experiments and results, followed by a conclusion in the last section.

2 Related Work/Literature

Multi label learning problem is generally realized by problem transformation and algorithm adaptation methods. Few popular algorithms under these categories are binary relevance method, label power set method, pruned sets method, C4.5, Adaboost.MH & Adaboost.MR, ML-kNN, Classifier chains method etc[20]. These methods either decomposes classification task into multiple independent binary classification tasks[6], one for each category or the ranking function of category labels from the labeled instances and apply it to classify each unknown test instance by choosing all the categories with the scores above the given threshold[20]. Almost all of these methods are supervised in nature. These methods cannot utilize information conveyed by unlabeled data. The other common drawbacks include inability to handle relationship among class labels and can not scale to large data set.

Recently some new approaches for multi-label learning that consider the correlations among categories have been developed. Few eg. are generative model proposed by Ueda[26], Bayesian model proposed by Griffiths [27], Hierarchical structure considered by Rousu [28], Maximum entropy method proposed by Zhu[29], Latent variable based approach proposed by McCallum. But all these methods are also supervised in nature.

Few recent approaches effectively used semi supervised learning for multi label text classification. In 2006 Liu, Jin and Yan proposed Multi-label classification approach based on constrained non negative matrix factorization [8]. In this approach parameter selection affects the overall performance of the system. Zha and Mie proposed Graph-based SSL for multi-label classification in the year 2008[9]. But this approach was purely intended for classification of video files and not for documents. Chen,Song and Zhang proposed Semi supervised multi-label learning by solving a Sylvester Eq in the year 2010 [10]. In this approach they constructed graph for input representation and class representation as well but this approach is getting slower on convergence when applied in the situation where large number of classes and input