Expert Systems With Applications 100 (2018) 157–164

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Hierarchy construction and text classification based on the relaxation strategy and least information model

Yongping Du[a,*], Jingxuan Liu[a], Weimao Ke[b], Xuemei Gong[b]

[a] Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
[b] College of Computing and Informatics, Drexel University, Philadelphia 19104, USA

ARTICLE INFO

Article history:
Received 28 November 2017
Revised 20 January 2018
Accepted 1 February 2018

Keywords:
Hierarchy classification
Relaxation strategy
Least Information Theory
Term weighting

ABSTRACT

Hierarchical classification is an effective approach to categorization of large-scale text data. We introduce a relaxed strategy into the traditional hierarchical classification method to improve the system performance. During the process of hierarchy structure construction, our method delays node judgment of the uncertain category until it can be classified clearly. This approach effectively alleviates the 'block' problem which transfers the classification error from the higher level to the lower level in the hierarchy structure. A new term weighting approach based on the Least Information Theory (LIT) is adopted for the hierarchy classification. It quantifies information in probability distribution changes and offers a new document representation model where the contribution of each term can be properly weighted. The experimental results show that the relaxation approach builds a more reasonable hierarchy and further improves classification performance. It also outperforms other classification methods such as SVM (Support Vector Machine) in terms of efficiency and the approach is more efficient for large-scale text classification tasks. Compared to the classic term weighting method TF*IDF, LIT-based methods achieves significant improvement on the classification performance.

1. Introduction

The task of text classification is to assign a predefined category to a free text document. With more and more textual information available online, hierarchical organization of text documents is becoming increasingly important to manage the data. The research on automatic classification of documents to the categories in the hierarchy is needed.

Most of the classifiers make the decision in the same flat space. Classification performance degrades quickly with larger scale data sets and more categories, especially in terms of the classification time. On the other hand, a hierarchical classification method organizes all of the categories into a tree like structure and trains a classifier on each node in the hierarchy. The classification process begins from the root of the tree until it reaches the leaf node which denotes the final category for the document.

The hierarchies are represented as binary trees mostly. During the hierarchical classification process, the document to be classified starts from the root and the next direction is determined by each node classifier. Finally, the leaf being reached will give the decision to its category label. However, there exists a 'blocking' problem during the process. The error that has occurred in the upper node classifier cannot be corrected by the lower node classifier. The 'blocking' problem may result in weaker performance compared to the non-hierarchical classification method. The advantage of hierarchy classification method is higher efficiency which is significant in large scale data set.

In order to improve the hierarchical classification performance, we introduce the relaxation strategy idea during the process of hierarchy construction and further propose the hierarchical classification approach based on it. The method delays the uncertain category decision until it can be classified definitely, thereby alleviating the impact of the 'blocking' problem. We give the experiment on the Reuters Corpus Volume 1(RCV1). The result denotes that our method can build a more rational category hierarchy and improve the performance of traditional hierarchy classification. Especially, the approach has higher time efficiency than other classifiers such as Support Vector Machine.

Another contribution of this work is in term weighting and documentation representation. The classic TF*IDF has been widely used for term weighting and document representation in text clustering and classification tasks (Liu, Liu, Chen, & Ma, 2003; Yang & Pedersen, 1997). Least Information Theory (LIT) extends Shannon's information theory to accommodate a non-linear relation between

* Corresponding author.
E-mail addresses: ypdu@bjut.edu.cn (Y. Du), Liujx99@emails.bjut.edu.cn (J. Liu), wk@drexel.edu (W. Ke), xuemeigong@drexel.edu (X. Gong).

https://doi.org/10.1016/j.eswa.2018.02.003
0957-4174/

information and uncertainty and offers a new way of modeling for term weighting and document representation (Ke, 2015). It establishes a new basic information quantity and provides insight into how terms can be weighted based on their probability distributions in documents vs. in the collection. We adopt the LIT for term weighting during hierarchical classification and it achieves significant performance improvement over classic TF*IDF.

2. Related work

It is important to build the rational category hierarchy and there are two common ways to implement this, including the Top-Down and Bottom-Up approaches. Liu, Yi, and Chia (2005) present a method to build up a hierarchical structure from the training dataset and uses the K-Means clustering algorithm to divide the category set. The hierarchical structure of the SVM classification tree manifests the interclass relationships among different classes.

Chen, Crawford, and Ghosh (2004) propose the segmentation approach using the maximum division strategy. It presents a new approach called HSVM (Hierarchical Support Vector Machines) to address multiclass problems. The method solves a series of max-cut problems to hierarchically and recursively partition the set of classes into two-subsets. The way of Bottom-Up cannot guarantee the separability of the category node set and the Top-Down approach is more commonly used.

Most hierarchy building methods organize the categories into a tree structure and usually the hierarchies are represented as binary trees which means that at each node a binary decision is made on which of the two subtrees to choose (Griffin & Perona, 2008). Marcin (2008) proposes a new idea which allows the child node has more than one parent node, and all the categories are organized as the DAG (Directed Acyclic Graph) structure. This idea has been used in the field of image classification and shows strong performance.

There are also some works that are focused on the Hierarchical Multi-label Classification. Each parent node is divided into multiple child nodes and the process is continued until each child node represents only one class. Zhang, Shah, and Kakadiaris (2017) consider the structural information embedded in the class hierarchy and uses it to improve the hierarchical classification performance. Bengio, Weston, and Grangier (2010) introduces an approach for fast multi-class classification by learning label embedding trees and it outperforms other tree-based or embedding approaches.

The traditional text classification approaches often require labeled data for learning classifiers, which is extremely expensive when applied to large-scale data involving thousands of categories. Viet (2011) takes advantage of the ontological knowledge for large-scale hierarchical text classification which does not require any labeled data. The classifier gets a reasonable performance. Pavlinek and Podgorelec (2017) presents the Self-Training LDA method for text classification in a semi-supervised manner with representations based on topic models.

The hierarchical classification approach decomposes the multi-class classification problem into different sub-task, and every node classifier solves the sub-task separately. The linear classifier (Deng, Satheesh, Berg, & Li, 2011), Bayesian Network (Wang, Wang, & Xie, 2011) and Support Vector Machine (Gao & Koller, 2011; Griffin & Perona, 2008) are used as the node classifier.

Another important aspect of this research is on feature selection and weighting for classification. In text clustering and classification research, TF*IDF has been extensively used for term weighting and document representation (Liu et al., 2003; Yang & Pedersen, 1997; Zhang, Wang, & Si, 2011). While term frequency (TF) indicates the degree of a document's association with a term, inverse document frequency (IDF) is the manifestation of a term's specificity, key to determine the term's value toward weighting and relevance ranking (Jones, 2004). Chen, Zhang, Long, and Zhang (2016) propose a new term weighting scheme TF-IGM (term frequency & inverse gravity moment) which incorporates a new statistical model to precisely measure the class distinguishing power of a term. Deepak, Kesari, and Priyanka (2017) propose a novel Variable Global Feature Selection Scheme (VGFSS) to select a variable number of features from each class based on the distribution of terms in the classes. While many classification algorithms have been developed, TF*IDF and its variations remain the de facto standard for term weighting in classification.

In IR (Information Retrieval), information and probability theories have provided important guidance to the development of classic techniques such as probabilistic retrieval and language modeling (Robertson & Zaragoza, 2009). The probabilistic retrieval framework provides an important theoretical ground to IDF weights (Robertson, 2004). IDF resembles the entropy formula in Shannon's information theory and several works have attempted to justify IDF from an information-theoretic view. IDF can be interpreted as Kullback–Leibler (KL) information (relative entropy) between term probability distributions in a document and in the collection (Aizawa, 2000). KL divergence measures information for discrimination between two probability distributions by quantifying the entropy change in a non-symmetric manner (Kullback & Leibler, 1951).

In the KL information view of IDF, the asymmetry of KL and infinite information it quantifies in special cases have undesirable consequences in the text classification context. From an information-centric view, Ke (2015) developed a new model for term weighting and document representation. By quantifying the amount of semantic information required to explain probability distribution changes, the proposed Least Information Theory (LIT) offers a new measure through which terms can be weighted based on their probability distributions in documents vs. in the collection. Several term weighting schemes such as LI Binary (LIB) and LI Frequency (LIF) were derived and experimented for text clustering. In this research, based on the notion of mutual information and the new LIT theory, we propose Least Information Gain for feature selection and combinations of other LIT-based methods for hierarchy construction and classification. We are interested in understanding the effectiveness of LIT in hierarchical classification tasks.

3. Hierarchy construction with relaxation strategy

3.1. Relaxation method

The category set will be divided into two subsets recursively to build a hierarchical structure that contains n categories. K-Means clustering algorithm is adopted to get the two clusters on the text data set, and it will help to determine which node the category belongs to.

As shown in Fig. 1, the aim is to divide the root node $\boldsymbol{S}$, which contains category ***A, B*** and $\boldsymbol{C}$, into two subsets referred to $\boldsymbol{S_L}$ and $\boldsymbol{S_R}$ respectively.

As an example, there are 30 training documents *A01,…, A10, B01,…, B10, C01,…, C10* in the root Node $\boldsymbol{S}$ and they belong to category ***A, B,*** and $\boldsymbol{C}$ respectively. These documents are clustered into two sets by K-Means with label $+1$ and -1. We find that most of the documents in category $\boldsymbol{A}$ are labeled as $+1$ and so the category $\boldsymbol{A}$ is assigned to $\boldsymbol{S_L}$. Similarly, the category $\boldsymbol{C}$ is assigned to $\boldsymbol{S_R}$. For category $\boldsymbol{B}$, there are 6 documents labeled as $+1$ and 4 documents labeled as -1. It is uncertain to decide which node the category $\boldsymbol{B}$ belongs to. We delay the decision to the next lower level by assigning the category $\boldsymbol{B}$ to $\boldsymbol{S_L}$ and $\boldsymbol{S_R}$ simultaneously. This relaxation idea will be used during the process of hierarchy construction.