MULTIVARIATE REGRESSION TREE TOPIC MODELING

Marco Ortu [1], Giulia Contu[1] and Luca Frigau[1]

[1] Dept. of Economics and Business Sciences, University of Cagliari, (e-mail: marco.ortu@unica.it, giulia.contu@unica.it, frigau@unica.it)

ABSTRACT: In this paper we propose Multivariate Tree Topic Modeling methodology, a general purpose approach to Topic Detection, which aims to refine the general results of a Topic Modeling methodology using Multivariate Trees in order to obtain consistent document groups. Topic modeling is defined as a mechanism for discovering low-dimensional, multi-faceted summaries of textual documents, typically by discovering hidden or latent topics in a corpus of documents. Given these hidden topics, we exploit the Multivariate Trees to obtain more homogeneous document groups with respect to the Topic Modeling output alone. We applied our model to a standard corpus of documents generally used in this kind of study to show that, when the aim of Topic Modeling is to generate coherent clusters of documents, the use of Multivariate Trees improves the overall coherence of these clusters for a wide range of Multivariate Trees' size.

KEYWORDS: Multivariate Analysis, Decision Trees, Topic Detection

1 Introduction

Topic modeling (TM) is a method for detecting latent structures in a collection of text documents. From the mathematical perspective, it can be seen as a dimensional reduction problem, where the vector space of a text document is often greater than several tens or hundreds of thousands, while the output vectorial space of topic modeling is typically in the order of tens and seldom hundreds. The typical use case of topic modeling is to represent themes and topics that are present in a large corpus of text data. This article proposes a method, based on Multivariate-Trees (MT), to refine the topic modeling output. In the literature, there are three main families of Topic Modeling methods: i) Matrix factorization based methods, ii) Probabilistic Methods, and more recently iii) Deep-learning based approaches. Matrix factorization-based topic modeling methods rely on a linear algebraic technique that factorizes a term-document matrix into two non-negative matrices, where one matrix represents

the topic-word distribution and the other matrix represents the document-topic distribution. The constraints used to solve the linear algebraic problem lead to the specific implementation, a widely used topic model is the Non-negative Matrix Factorization (NMF) (Lee & Seung, 2000). Probabilistic-based topic modeling methods rely on the hypothesis that documents are a mixture of topics guided by, typically, two hidden distributions of words in a collection of documents, one that models the distribution of words in hidden topics and one that models the distribution of topics in documents. One of the most popular approaches to topic modeling is Latent Dirichlet Allocation (Blei *et al.*, 2003) (LDA). It is a generative probabilistic model that assumes that each document in a corpus is generated by a mixture of latent topics, where a distribution over words characterizes each topic. When used for clustering, documents are grouped together by their dominant topic. However, these groups might be incoherent, since the assignment of the dominant topic is arbitrary. We show that, when the main goal of topic modeling is to generate a coherent clustering of documents, the use of multivariate trees improves the overall coherence of these clusters as measured by heterogeneous indexes such as Gini's Index.

2 Methodology And Data

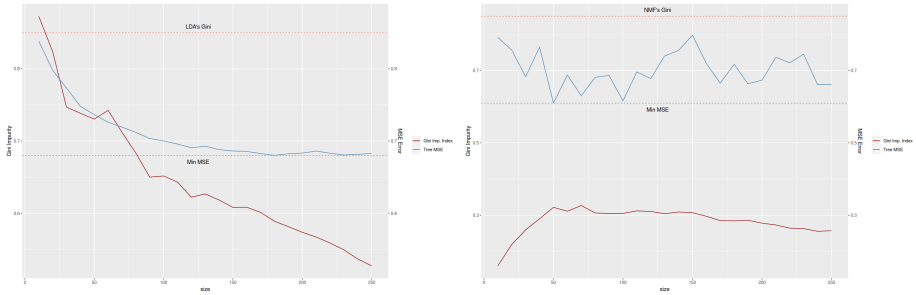
Multivariate-Tree Topic Modelling De'Ath, 2002 is a general purpose approach to Topic Modeling, which aims to refine the general results of a TM methodology using Multivariate Trees in order to obtain consistent documents groups. Algorithm 1 illustrates MTTM method. It is composed of two phases. In the first phase, a topic model is used to obtain the topic distributions of each document. The topic model leads to a topic distribution over the documents, as often in these applications, we considered the dominant topic, namely ϕ_t, and grouped the documents according to it. We finally evaluate the average Gini's index of each group considering the true category, namely y_t, obtaining our baseline measure of the groups' homogeneity ($\bar{G}_{topic}$). Our goal is to maximize the homogeneity of the groups (thus minimizing the average Gini's index). In the second phase, we apply a multivariate regression tree using the topic distributions as dependent variables and the words' frequencies as predictors. We evaluate the average Gini's index $\bar{G}_{tree}$ as a function of tree's size, and compare it with the topic baseline ($\bar{G}_{topic}$).

Figure 1 shows the results of MTTM for the 20 Newsgroups dataset using MT (De'Ath, 2002). Figure 1a reports our results using LDA topic modeling. The dashed red line is the baseline of the Gini's index obtained by the topic modeling alone, it is the average Gini's index of the topic modeling groups

Algorithm 1 MTTM Algorithm Definition

Require $\mathcal{D}$: Training documents of size N, each with a categorical response variable y_t and a set of quantitative variables X_t;
Step 1:
Input: $\mathcal{D} = \{d_1, \ldots, d_N\}$
Output: $\phi : \mathcal{W} \rightarrow \Theta$
$X \leftarrow W$
$Y \leftarrow \phi$
$\bar{G}_{topic} = \frac{1}{T} \sum_{i=1}^{T} G_t(\phi_t)$
Step 2:
set $(s_{min}; s_{max})$
for $s = s_{min}$, $s{+}{+}$, $s < s_{max}$ **do**
fit MT: $Y \sim X$
$\bar{G}_{tree} = \frac{1}{s} \sum_{i=1}^{s} G_s(y_t)$
end for

considering the true category. The green dashed line represents the Gini's index of the groups obtained by the multivariate tree obtained for the tree size that minimizes the tree MSE. The blue continuous line represents the tree MSE over the tree size, and the red continuous line represents the average Gini's Index over the tree size. We can observe that the average Gini's Index of the multivariate tree decreases as the number of splits increases, at the limit case when the number of splits equals the number of documents, each split contains only one document and the Gini's index is zero. In this particular case, we knew that there were 20 topics, thus we run the LDA algorithm using 20 as the number of topics. Analyzing the output of LDA, we could identify only nine out of 20 topics, namely, only nine topics were assigned with a probability greater than zero. On the other hand, the multivariate tree's output showed an optimal number of splits around about 150. In this range of the number of splits, we can observe that the MTTM output always yields more homogeneous groups. Figure 1b shows the results using NMF topic modelling. In this case, it can be observed that the average Gini's Index over the tree size exhibit a different behaviour, it starts with a lower value of average Gini's index, and it increases as the tree's size increases, then it stabilizes for a wide range of tree's size and finally start a decreasing trend for extreme values. In general, from Figure 1 we can observe that, after the application of the MT, the overall Gini's index is improved for a wide range of the tree's size.



(a) LDA algorithm. (b) NMF algorithm.

Figure 1: Average Gini's impurity index for the 20 Newsgroups dataset using Multivariate Tree.

3 Conclusion

This study proposes a methodology to enhance the ability of a topic detection method to create coherent groups leveraging decision trees. We presented the Multivariate-Tree Topic Modelling framework MTTM. MTTM is constructed by combining topic modeling and multivariate-trees methodologies. We applied our model to a standard corpus of documents (the 20 Newsgroup) and two topics models (LDA and NMF) generally used in this kind of studies, and we found that, when the aim of TM is to generate coherent clusters of documents, the use of a MT improves the overall coherence of these clusters for a wide range of the MT's size.

References

BLEI, DAVID M, NG, ANDREW Y, & JORDAN, MICHAEL I. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, **3**, 993–1022.

DE'ATH, GLENN. 2002. Multivariate regression trees: a new technique for modeling species–environment relationships. *Ecology*, **83**(4), 1105–1117.

LEE, DANIEL, & SEUNG, H SEBASTIAN. 2000. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, **13**.