Expert Systems With Applications 91 (2018) 159–169

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Stability of topic modeling via matrix factorization

Mark Belford*, Brian Mac Namee, Derek Greene

Insight Centre for Data Analytics, University College Dublin, Ireland

ARTICLE INFO

Article history:
Received 14 February 2017
Revised 11 August 2017
Accepted 28 August 2017
Available online 1 September 2017

Keywords:
Topic modeling
Topic stability
LDA
NMF

ABSTRACT

Topic models can provide us with an insight into the underlying latent structure of a large corpus of documents. A range of methods have been proposed in the literature, including probabilistic topic models and techniques based on matrix factorization. However, in both cases, standard implementations rely on stochastic elements in their initialization phase, which can potentially lead to different results being generated on the same corpus when using the same parameter values. This corresponds to the concept of "instability" which has previously been studied in the context of *k*-means clustering. In many applications of topic modeling, this problem of instability is not considered and topic models are treated as being definitive, even though the results may change considerably if the initialization process is altered. In this paper we demonstrate the inherent instability of popular topic modeling approaches, using a number of new measures to assess stability. To address this issue in the context of matrix factorization for topic modeling, we propose the use of ensemble learning strategies. Based on experiments performed on annotated text corpora, we show that a K-Fold ensemble strategy, combining both ensembles and structured initialization, can significantly reduce instability, while simultaneously yielding more accurate topic models.

1. Introduction

Topic models aim to discover the latent semantic structure or topics within a corpus of documents, which can be derived from co-occurrences of words across the documents. Popular approaches for topic modeling have involved the application of probabilistic algorithms such as Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003). More recently, Non-negative Matrix Factorization approaches (Lee & Seung, 1999) have also been successfully applied to identify topics in unstructured text (Arora, Ge, & Moitra, 2012; Kuang, Choo, & Park, 2015).

The standard formulations of both the LDA and NMF algorithms include stochastic elements in their initialization phase, prior to an optimization phase which produces a local solution. This random component can affect the final composition of the topics found and the rankings of the terms that describe those topics. This is problematic when seeking to capture a definitive topic modeling solution for a given corpus and represents a fundamental *instability* in these algorithms – different runs of the same algorithm on the same data can produce different outcomes. This problem has been widely studied in the context of partitional clustering algorithms such as *k*-means, which tends to converge to one of numerous local minima, depending on the choice of starting condition (Bradley & Fayyad, 1998). It has long been recognized as a significant drawback of such algorithms and a substantial number of works exist which attempt to address the issue (*e.g.* Kuncheva & Vetrov, 2006; Pena, Lozano, & Larranaga, 1999).

In the case of topic modeling, instability can manifest itself in two distinct aspects. The first can be observed when examining the topic descriptors (*i.e.* the top terms representing each topic) over multiple runs. The term rankings may change considerably, where certain terms may appear or disappear completely between runs. Secondly, issues of instability can also be observed when examining the degree to which documents have been associated with topics across different runs of the same algorithm on the same corpus. In both cases, such inconsistencies can potentially alter our interpretation and perception of a given topic model. Also, it is clear that any individual run should not be treated as a "definitive" summary of the underlying topics present in the data.

Generally speaking, in the comparative evaluation of topic modeling approaches, researchers tend to focus on either the coherence of the topic descriptors (Newman, Lau, Grieser, & Baldwin, 2010) or the extent to which the topics accurately coincide with a set of ground truth categories or human annotations (Kuang et al., 2015). However, few researchers have considered the evaluation of different approaches from the point of view of their stability across multiple runs.

* Corresponding author.
E-mail addresses: mark.belford@insight-centre.org (M. Belford), brian.macnamee@ucd.ie (B. Mac Namee), derek.greene@ucd.ie (D. Greene).

http://dx.doi.org/10.1016/j.eswa.2017.08.047
0957-4174/

In this paper we quantitatively assess the extent to which standard randomly-initialized NMF and LDA algorithms are unstable with respect to the topics that they produce on a diverse collection of text corpora. To do this we propose measures that capture the two distinct aspects of instability outlined above. We then focus on addressing the issue in the context of matrix factorization, exploring the use of strategies that involve improved initialization and ensemble learning. In particular, we propose a new combined approach, motivated by the traditional concept of k-fold cross-validation, which can yield stable results while also often producing more accurate and coherent models.[1]

The rest of the paper is structured as follows. In Section 2 we provide an overview of relevant work in topic modeling and the more general area of cluster analysis. In Section 3 we discuss the problem of topic model instability in more detail, describing three new measures to quantify instability in topic models. In Section 4 we propose ensemble approaches to address the issue, which are subsequently evaluated on ten different text corpora in Section 5. Finally in Section 6 we conclude the paper with ideas for future work.

2. Related work

2.1. Topic modeling

Topic models attempt to discover the hidden thematic structure within an unstructured collection of text without relying on any form of training data. These models date back to the early work on latent semantic analysis (LSA) by Deerwester, Dumais, Landauer, Furnas, and Harshman (1990), who proposed applying SVD to decompose a document-term matrix to uncover the associations between terms and concepts in the data. In basic terms, a topic model consists of k topics, each represented by a ranked list of strongly-associated terms (often referred to as a "topic descriptor"). Each document in the corpus can also be associated with one or more of these topics to varying degrees.

Considerable research on topic modeling has focused on the use of probabilistic methods, where a topic is viewed as a probability distribution over words, with documents being mixtures of topics (Steyvers & Griffiths, 2007). The most widely-applied probabilistic topic modeling approach has been LDA (Blei et al., 2003). Different approximation methods have been proposed for LDA inference, including variational inference and Markov chain Monte Carlo (MCMC). Such approximation algorithms can converge to different local maxima on the same data (Zhao et al., 2015). The most commonly-used implementation, provided by the *Mallet* software package (McCallum, 2002), relies on fast Gibbs sampling, where the initial state is determined by a user-specified random seed.

Alternative algorithms, such as Non-negative Matrix Factorization (Lee & Seung, 1999), have also been effective in discovering topics in text corpora (Arora et al., 2012; Kuang et al., 2015). NMF is an unsupervised approach for reducing the dimensionality of non-negative matrices. When working with a document-term matrix $\mathbf{A}$, the goal of NMF is to approximate this matrix as the product of two non-negative factors $\mathbf{W}$ and $\mathbf{H}$, each with k dimensions. The rows of the factor $\mathbf{H}$ can be interpreted as k topics, defined by non-negative weights for each of the m terms in the corpus vocabulary. Ordering each row provides a topic descriptor, in the form of a ranking of the terms relative to the corresponding topic. The columns in the matrix $\mathbf{W}$ provide membership weights for all documents with respect to each of the k topics. One of the advantages of NMF over traditional LDA methods is that there are fewer parameter choices involved in the modeling process, while it also has a tendency to identify more coherent topics than LDA (O'Callaghan, Greene, Carthy, & Cunningham, 2015).

NMF is commonly initialized by assigning random non-negative weights to the entries in the factors $\mathbf{W}$ and $\mathbf{H}$. By applying an optimization process, such as alternating least squares (Lin, 2007), the factors are iteratively improved to reduce the approximation error until a local minimum is reached. As a result, the values in the initial pair of factors will have a significant impact on the values in the final factors (*i.e.* the topic-term and topic-document weights), even after a large number of iterations have been performed. Alternative initialization schemes for NMF have focused on increasing the accuracy of the final factors by using a more structured process, such as seeding using a prior clustering algorithm (Wild, Curry, & Dougherty, 2004). Another approach, Non-negative Double Singular Value Decomposition (NNDSVD) (Boutsidis & Gallopoulos, 2008), chooses initial factors based on a sparse SVD approximation of the original data matrix. This has been shown to be particularly effective on sparse data, such as text (O'Callaghan et al., 2015). In its basic form, NNDSVD contains no stochastic element and should technically converge to the same pair of factors each time, although this depends on the underlying SVD implementation being used.

2.2. Stability in cluster analysis

Partitional clustering algorithms, such as k-means and k-medoids, have an inherent stability problem. That is to say, if we run the same algorithm on the same data or data drawn from the same source repeatedly, we frequently achieve different results between each run. This variation can either be due to poor random seeds leading to convergence to different local minima (Pena et al., 1999), or as a result of perturbations in the data (Ben-Hur, Elisseeff, & Guyon, 2002).

One widely-adopted approach for dealing with the issue is to adopt a better cluster initialization strategy that is either fully deterministic or at least produces less variation than random initialization, while simultaneously yielding more useful clusterings. A popular initialization approach proposed by Arthur and Vassilvitskii (2007), referred to as *k-means++*, involves choosing an initial seed item at random as the first cluster center and then choosing each subsequent cluster center with a probability proportional to its squared distance from the items nearest existing cluster centers. To further improve the resulting clustering, this process can be repeated for several different initial seed items. While this strategy is not deterministic, it does tend to yield more consistent results across multiple runs. Researchers have also proposed fully deterministic strategies, where initial cluster centers are determined based on embedding methods such as PCA (Su & Dy, 2004), or coming from the prior application of another algorithm such as hierarchical clustering (Celebi & Kingravi, 2012).

2.3. Ensemble methods

An alternative strategy for reducing instability in unsupervised learning is to use ensemble clustering techniques, which are based on the premise that combining large, diverse sets of clusterings can produce a more stable and accurate solution (Strehl & Ghosh, 2002). Ensemble approaches are usually divided into two different stages. Firstly, a collection of base clusterings are generated (*i.e.* the ensemble members), typically by repeatedly applying an algorithm such as k-means with random initialization to the full dataset or to random samples of the data (Minaei-Bidgoli, Topchy, & Punch, 2004). Secondly, an integration function is applied to combine the base clusterings into a single consensus clustering. One of the most common integration strategies utilized is to leverage information from the ensemble regarding the

[1] See https://github.com/derekgreene/topic-ensemble.