An Efficient Nonnegative Matrix Factorization Topic Modeling for Business Intelligence

[1]K PrashantGokul, [2]M.Sundararajan
{[1]kprashantgokul@gmail.com ,[2]msrajan69@gmail.com}

[1]Research Scholar, Dept. of ECE, Bharath Institute of Higher Education and Research, Chennai, India. [2]Professor & Pro-VC, Bharath Institute of Higher Education and Research, Chennai, India.

Abstract. Topic models can give us a knowledge into the basic latent design of an enormous corpus of documents. A scope of strategies have been planned in the writing, including probabilistic topic models and methods dependent on matrix factorization. Notwithstanding, the subsequent topics frequently address just broad, in this manner excess information about the data instead of minor, yet possibly significant information to clients. To handle this issue, we propose a novel sparseness improvement model of negative matrix factorization for finding excellent nearby topics. In any case, in the two cases, standard executions depend on stochastic components in their instatement stage, which can possibly prompt various outcomes being produced on a similar corpus when utilizing a similar boundary values. To address this issue in the context of matrix factorization for topic modeling, we propose the utilization of ensemble learning procedures. We show the useful utility of ENMF on New York Times dataset, and find that ENMF is particularly helpful for applied or expansive topics, where topic key terms are not surely known. We find that ENMF accomplishes higher weighted Jaccard similarity scores than the contemporary strategies..

Keywords: Topic modeling, Factorization, Ensemble, Clustering.

1 Introduction

As our collection of computerized documents keeps on being put away and gets immense, we just don't have the human ability to peruse the entirety of the documents to give topical data. Therefore, we need customized instruments for removing the effective data from the assortment. Theme displaying is such a factual model that has been shown viable for this endeavor including discovering subjects and their examples as time goes on. Theme displaying is an unaided learning as in it needn't waste time with names of the archives. The points are mined from text based substance of the archives. All things considered, the general issue for point displaying is to use the saw archives to find the mysterious subject designs. Furthermore, with the discovered themes we can organize the assortment for certain reasons, for instance requesting, overview, dimensional reduction [1] , Latent Dirichlet portion (LDA) [3] is a renowned probabilistic point model. It was made to fix a couple of issues with an in the past developed theme model probabilistic dormant semantic investigation (pLSA) [6]. LDA expects that a report typically addresses various points which are demonstrated as spread over a language. Each word in the record is made by indiscriminately picking a point from an appointment over subjects, and a while later discretionarily picking a word from a spread over the language. The normal procedures to enlist back of the model are harsh allowance

I3CAC 2021, June 07-08, Chennai, India

DOI 10.4108/eai.7-6-2021.2308681

techniques. Incredibly, the most limit probability approximations are NP-hard. Along these lines, a couple of specialists continue arranging calculations with provable confirmations for the issue of learning the subject models. These calculations consolidate nonnegative network factorization (NMF) [2]. In various applications, the records may contain metadata that we ought to fuse into subject demonstrating. Titles and labels are occurrences of the metadata that by and large go with the records in various applications. This metadata is truly made by human to depict the effective data of archives. It gets huge considering the way that reflects the essential subjects of reports just as has a limited construction. In like manner, a fitting technique to consolidate this metadata to theme displaying is depended upon to improve the show of point demonstrating. Clearly, the techniques that address the issue of fusing these metadata into NMF-based point models are at this point phenomenal. The fundamental method to manage fuse the metadata into NMF-based point demonstrating is by uniting the metadata and the text based substance of records, and a while later separating themes from this affiliation set. The relationship of both text based informational substance structures are found by a NMF calculation from tags. Having these topic-element structures, the separated topics are enhanced by words existing in textual substance identified with the element utilizing a NLS calculation at more significant level. As of late, a strategy called nonnegative various matrix factorization is proposed [11]. This technique incorporates the metadata as an assistant matrix that imparts segment to the substance matrix and afterward decays the two matrices at the same time. According to specialized perspective, this strategy is like OLLH which extricates topics from the substance and the metadata together. In addition, this strategy is material just for a particular NMF calculation.

In this paper First the technique is reached out to be appropriate for overall NMF algorithm. At the inferior level, topics is found by a NMF calculation substance. Given the topics and the substance, topic-content constructions are approximated utilizing a NLS calculation. Having these topic-content designs, the separated topics are improved by words existing in the metadata utilizing a NLS calculation at more significant level. For instance, some online news entries share total titles and just little piece of substance, yet different applications may share the two titles and substance in a total structure. Besides, the analyses show that TLLH isn't just more proficient yet it additionally gives higher interoperability scores than OLLH. The patterns of removed principle topics throughout a time-frame might be utilized as foundation information for different applications, for example sentiment analysis [13].

Notwithstanding, when practical in visual examination, LDA has a few reasonable inadequacies in term of constancy from numerous runs and exact assembly. Besides, because of the complicatedness in the detailing and the calculation, incorporating different sorts of client criticism with LDA is relatively difficult. As a methodology for topic modeling, NMF works like LSI in that the two of them tackle a matrix deterioration issue given a specific position esteem relating to the quantity of topics. Nonetheless, as the name recommends, NMF forces non-antagonism requirements on each component of the subsequent matrices with the goal that it can look after interpret-ability. In addition, the NMF calculation is deterministic. Subsequently, except if the client modifies an underlying specification, she will acquire a similar outcome from the calculation. These attractive practices of NMF fill in as significant grounds to make UTOPIAN basically valuable and intelligent in true visual investigation by empowering the client to dynamically improve a specific outcome by intuitively changing the calculation specifications. The way wherein the embraced semisupervised NMF strategy

considers the client intercessions is instinctive on the grounds that the semi-management will be in a similar structure as the two above-portrayed topic modeling yields which the client is now acquainted with all through his/her analysis. This trademark eliminates any extra requirement for changing the client mediations back to the calculation boundaries or limitations in a vague manner.

2 RELATED WORK

Therefore, we need customized instruments for removing the effective data from the assortment. Theme displaying is such a factual model that has been shown viable for this endeavor including discovering subjects and their examples as time goes on. Theme displaying is an unaided learning as in it needn't waste time with names of the archives. The points are mined from text based substance of the archives. All things considered, the general issue for point displaying is to use the saw archives to find the mysterious subject designs. Furthermore, with the discovered themes we can organize the assortment for certain reasons, for instance requesting, overview, dimensional reduction [1] , Latent Dirichlet portion (LDA) [3] is a renowned probabilistic point model. It was made to fix a couple of issues with an in the past developed theme model probabilistic dormant semantic investigation (pLSA) [6]. LDA expects that a report typically addresses various points which are demonstrated as spread over a language. Titles and labels are occurrences of the metadata that by and large go with the records in various applications. This metadata is truly made by human to depict the effective data of archives. It gets huge considering the way that reflects the essential subjects of reports just as has a limited construction. In like manner, a fitting technique to consolidate this metadata to theme displaying is depended upon to improve the show of point demonstrating. Clearly, the techniques that address the issue of fusing these metadata into NMF-based point models are at this point phenomenal. The fundamental method to manage fuse the metadata into NMF-based point demonstrating is by uniting the metadata and the text based substance of records, and a while later separating themes from this affiliation set.

3 An Ensemble approach for NON NEGATIVE MATRIX FACTORIZATION (ENMF)

3.1 Sparseness

The idea of thin coding alludes to an illustrative plan where a couple of units are viably used to address ordinary data vectors. In actuality, this infers most units taking values near nothing while just scarcely any take significantly non-zero values. Various sparseness events have been future and utilized in the writing to date. Such measures are mappings from Rn to R which evaluate how much energy of a vector is pressed into a couple of parts. On a standardized scale, the sparsest conceivable vector ought to have a sparseness of one, though a vector with all components equivalent ought to have a sparseness of nothing. We utilize a thinnessamount dependent on the connection among the L1 standard and the L2 standard:

Sparseness (x) = √n−(∑|xi|)/q∑x2 i √n−1 (3.1)

where n is the dimensionality of x. This capacity assesses to solidarity if and just if x contain just a solitary non-zero part, and takes an estimation of nothing if and just if all segments are equivalent, introducing easily between the two limits. Our point is to oblige NMF to find arrangements with wanted levels of that point: what precisely ought to be sparse? The premise vectors W or the coefficients H? This is an inquiry that can't be offered an overall response; everything relies upon the specific application being referred to. Further, simply rendering the data matrix switches the job of the two, so it is not difficult to see that the decision of which to compel should be made by the experimenter. For instance, a specialist dissecting infection patterns may expect that most sicknesses are uncommon yet that every illness can cause an enormous number of manifestations.

Algorithm: NMF with sparseness constraints

1. Prepare W and H to irregular positive matrice

2. If sparseness limitations on W smear, at that point project every segment of W to be non-negative, have unaltered L2 standard, however L1 standard set to accomplish wanted sparseness

3. If sparseness imperatives on H apply, at that point project each column of H to be non-negative, have unit L2 standard, and L1 standard set to accomplish wanted sparseness

4. Iterate

    a. If sparseness imperatives on W apply,

W := W−μW(WH−V)HT
W := W⊗(VHT) (WHHT)

    b. If sparseness imperatives on H apply,

H := H−μHWT (WH−V)

H := H⊗(WTV) (WTWH)

Above, and mean elementwise duplication and division, separately. Additionally, W and H are little positive constant which should be set fittingly for the calculation to work. Luckily, they client; our execution of the calculation naturally adjusts these boundaries.

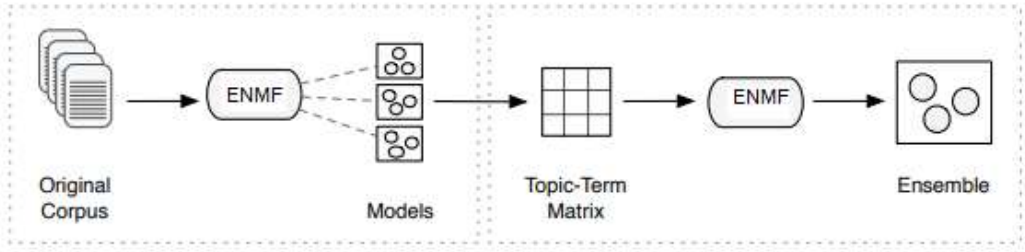

Fig 3.1 ENMF Architecture

We apply group learning for point displaying as two layers of network factorization. Fig. 3 shows a framework of the strategy, which can normally be disengaged into two phases, such as current systems in troupe grouping . Age: Create a lot of base theme models by executing r runs of NMF applied to a comparative corpus, tended to as a report term network A. Compromise:.

4.Experimental Setup

We utilize a solitary rundown of basic English stop-words for all datasets. ENMF works on bag-of-words text portrayals, as was applied to the crude frequency values. In our investigations, we think about two distinctive topic modeling draws near:

1. Standard NMF, NMF is usually introduced by allotting irregular nonnegative loads to the sections in the variables. By applying an advancement process, like rotating least squares, the elements are iteratively improved to decrease the guess blunder until a neighborhood least is reached. Therefore, the values in the underlying pair of variables will altogether affect the values in the last factors, even after countless emphasess have been performed [14].

4.1 Evaluation of Topic Models:

This segment presents the outcomes from the quantitative assessment of the ENMF, NMFD, NNDSVD algorithms by the Cv, UMass, and RS metrics. To start with, the outcomes from the assessment of ENMF, NMFD, NNDSVD algorithms models learned by the Newyorktimes data are introduced. The two ideal model's comparing algorithms from this assessment are then applied to assess how these algorithms sum up when learned with more heterogeneous data. The ENMF, NMFD, NNDSVD algorithms results are introduced in Table 6.1-Table 6.3 and Figure 6.2, Figure 6.3 and Figure 6.4,. A reasonable declining pattern was noticed for every one of the four learned ENMF models as the quantity of topics K expanded. The Cv pattern was not monotonic in its reduction nonetheless and had nearby variety spikes for little scopes of K, showing that there can be a neighborhood ideal of various topics to find in little neighborhood ranges of K.

Table 4.1 Cv coherence

Corpus	ENMF	NNDSVD	NMF
10	0.79	0.75	0.7
20	0.78	0.72	0.69
30	0.77	0.7	0.65
40	0.73	0.69	0.63
50	0.71	0.68	0.6
60	0.68	0.64	0.59
70	0.67	0.62	0.58
80	0.63	0.6	0.55
90	0.62	0.59	0.54
100	0.6	0.58	0.51

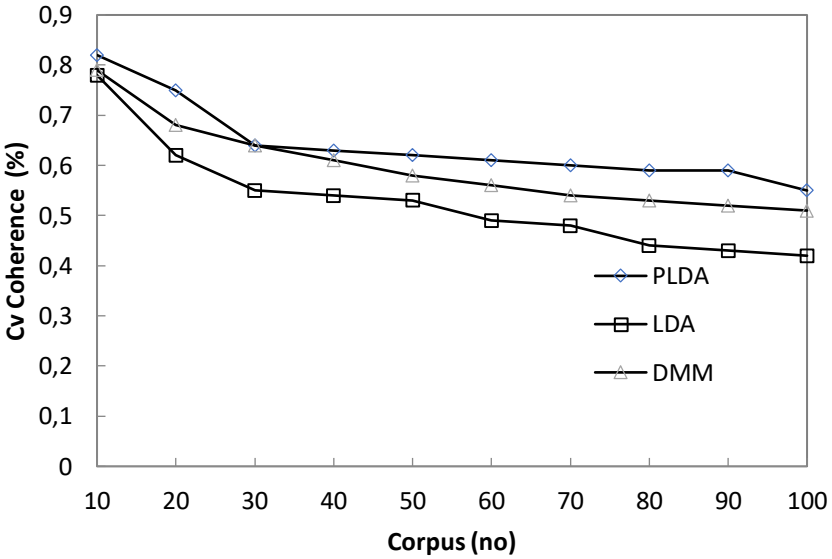

Fig .4.2 Cv Coherence

This declining trend was also observed for NMF, especially in UMass and RS, but it was not as pronounced as for ENMF. When comparing the ENMF and NNDSVD, NMFmodel it was even more clear, compared to the general case, that the optimal NNDSVD, NM did not have this notable declining trend. The ENMF achieved stable coherence scores for any number of topics up to 100 but had slightly decreasing coherence scores from 10 to 100 topics.

Table 4.2 UMass

	ENMF	NNDSVD	NMF
10	-3.2	-2.2	-1.3

20	-3.3	-2.3	-1.6
30	-3.8	-2.8	-1.9
40	-3.9	-2.9	-2.3
50	-4.2	-3.1	-2.5
60	-4.3	-3.3	-2.9
70	-4.6	-3.6	-3.2
80	-5.1	-3.7	-3.3
90	-5.2	-3.9	-3.6
100	-5.6	-4	-3.7

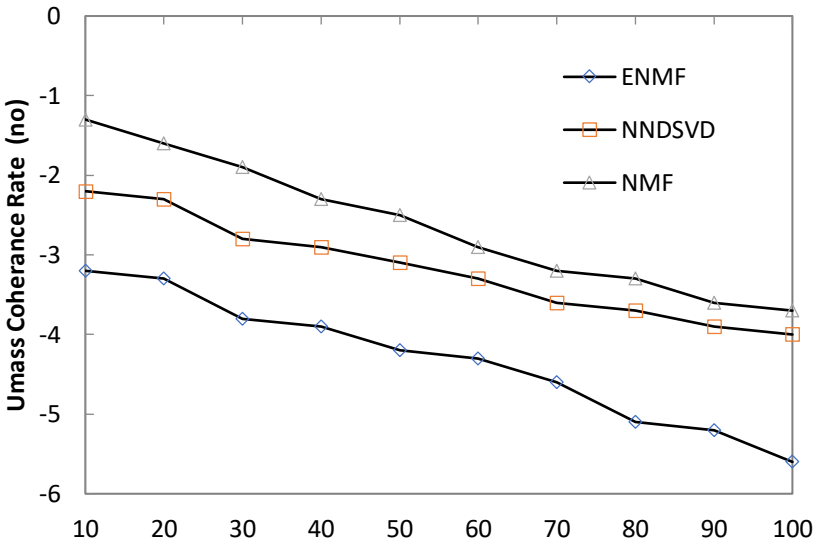

Fig .4.3 Umass Coherence

Although the ENMF model was superior in the Cv coherence score to the NMF, NNDSVD model for the lower end of the topic spectrum, they both had approximately the same coherence score for the higher end of the spectrum. For UMass these two versions were alike. The ENMF did perform better in the bottom 10 percent aggregation for both coherence scores, however, and in particular for UMass compared to the NMFD and NNDSVD. For the RS score, the ENMF exceeded the optimal NMF on average, while the bottom 10 percentile was alike.

Table 4.3 RS Score

No.of Corpus	**ENMF**	**NNDSVD**	**NMF**
10	0.45	0.4	0.38

20	0.48	0.41	0.39
30	0.52	0.42	0.41
40	0.56	0.45	0.42
50	0.61	0.49	0.45
60	0.62	0.51	0.47
70	0.63	0.53	0.52
80	0.67	0.54	0.59
90	0.7	0.57	0.61
100	0.71	0.61	0.62

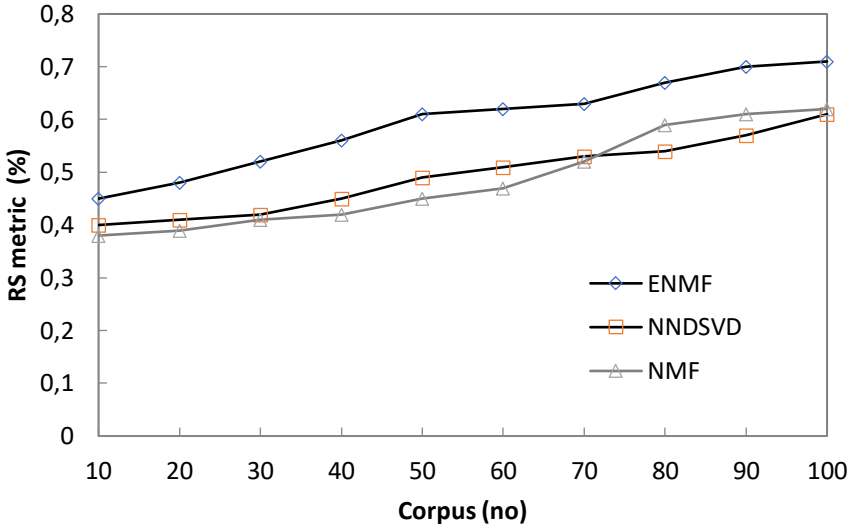

Fig .4.4 RS Metric

5. Conclusion

In these articles we introduced a novel subject demonstrating method, ENMF, that improves human interpretability of themes found from gigantic, deficiently understood corpora of reports. In such applications, ENMF engages the customer to fuse oversight by giving examples of archives needed point structure. Notwithstanding the way that we focused in on the enumerating of ENMF for subject demonstrating, this procedure can, with fitting choice of oversight, be instantly summarized to any non-negative framework decay application. We developed an iterative estimation for ENMF subject to multiplicative updates and exhibited the monotonic it of the computation and its association to a local ideal. Finally, we ran ENMF on the New York Times stood out the method from bleeding edge subject displaying techniques. We have shown that ENMF is an effective subject displaying methodology that should be considered in applications when human interpretability is critical.

References

[1] H Wu and Z Liu. “Non-negative matrix factorization with constraints”, In Proceedings of the 24th AAAI Conference on Artificial Intelligence, pp. 506–511, 2010.

[2] Gen Li, Dan Yang, Andrew B Nobel, and HaipengShen. “Supervised singular value decomposition and its asymptotic properties”, Journal of Multivariate Analysis, vol.146, pp.7–17, 2016.

[3] Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov. “Metagenesandmolecular pattern discovery using matrix factorization”, Proceedings of the national academyof sciences, vol.101, no.12, pp.4164–4169, 2004.

[4] Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. “Read ing tea leaves: How humans interpret topic models”, In Neural Information Processing Systems,2009.

[5] Yanhua Chen, ManjeetRege, Ming Dong, and Jing Hua. “Non-negative matrix factorization for semi-supervised data clustering”, Knowledge and Information Systems, vol.17, no.3, pp.355–379, 2008.

[6] Steven Bird, Ewan Klein, and Edward Loper. “Natural Language Processing with Python”, O’Reilly Media, Inc., 1st edition, 2009.

[7] David M Blei. “Probabilistic topic models”, Communications of the ACM, vol.55, no.4, pp.77–84, 2012.

[8] S. Arora, R. Ge, Y. Halpern, D. Mimno, and A. Moitra. “A practical algorithm for topic modeling with provable guarantees”, In proceeding of the 30th International Conference on Machine Learning, 2013.

[9] S. Arora, R. Ge, and A. Moitra, “ Learning topic models-going beyond svd”, In Proceeding of the IEEE 53rd Annual Symposium on Foundations of Computer Science, pages 1–10, 2012.

[10] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J, “Plemmons. Algorithms and applications for approximate nonnegative matrix factorization”, Computational Statistics and Data Analysis, vol.15, no.1, pp.155–173, 2007.

[11] V. Bittorf, B. Recht, C. Re, and J. A. “Factoring nonnegative matrices with linear programs”, In Neural Information Processing Systems, 2012.

[12] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, “New avenues in opinion mining and sentiment analysis”, IEEE Intelligent Systems, vol.28, no.2, pp.15–21, 2013.

[13] R. Feldman. Techniques and applications for sentiment analysis. Communication of the ACM, vol.56, no.4, pp.82–89, 2013.

[14] Lee, D.D.: “Learning the parts of objects by non-negative matrix factorization. Nature” vol.401, no.6755, pp.788–791, 1999

[15] Gonzalez, E.F., Zhang, Y, “Accelerating the lee-seung algorithm for nonnegative matrix factorization”, Department of Computational Applied Mathematics Rice University(CAAM) Houston TX Technical, pp. 1–13, 2005.

[16] Ding, C.H.Q., Tao, L., Jordan, “M.I.: Convex and semi-nonnegative matrix factorizations”. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.32, no.1, pp.45–55, 2010.

[17] M. Berry, M. Browne, A. Langville, V. Pauca and R. Plemmons, "Algorithms and Applications for Approximate Nonnegative Matrix Factorization", *Computational Statistics and Data Analysis*, vol. 52, no. 1, pp. 155-173, 2007.

[18] M. Gupta and J. Xiao, "Non-Negative Matrix Factorization as a Feature Selection Tool for Maximum Margin Classifiers", *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011.

[19] Z. Yang, H. Zhang, Z. Yuan and E. Oja, "Kullback-Leibler Divergence for Nonnegative Matrix Factorization", *Artificial Neural Networks and Machine Learning*, vol. 6791, pp. 250-257, 2011