COVID-19 or Flu? Discriminative Knowledge Discovery of COVID-19 Symptoms from Google Trends Data

Md Imrul Kaish
University of Texas Rio Grande Valley
mdimrul.kaish01@utrgv.edu

Md Jakir Hossain
University of Texas Rio Grande Valley
mdjakir.hossain01@utrgv.edu

Evangelos E. Papalexakis
University of California Riverside
epapalex@cs.ucr.edu

Jia Chen
University of California Riverside
jia.chen81uta@gmail.com

ABSTRACT

Google Trends data analytics is gaining more attention in the past few years, and most of the state-of-the-art algorithms are focused on forecasting. How to extract knowledge about symptoms mostly related to COVID-19 by contrasting periods of time with and without the spread of COVID-19 from Google Trends data has not been investigated. To this end, we propose a novel nonnegative discriminative analysis (DNA) to extract the unique information of one dataset relative to another dataset. Numerical tests corroborated the efficacy of our proposed approaches to discover the three unique COVID-19 symptoms w.r.t. flu including *ageusia*, *shortness of breath*, and *anosmia* while prior arts are not able to.

KEYWORDS

Nonnegative matrix factorization, Google Trends data, discriminative dimensionality reduction, COVID-19 symptoms

ACM Reference Format:
Md Imrul Kaish, Md Jakir Hossain, Evangelos E. Papalexakis, and Jia Chen. 2021. COVID-19 or Flu? Discriminative Knowledge Discovery of COVID-19 Symptoms from Google Trends Data. In *epiDAMIK 2021: 4th epiDAMIK ACM SIGKDD International Workshop on Epidemiology meets Data Mining and Knowledge Discovery*. ACM, New York, NY, USA, 3 pages. https://doi.org/xx.xxxx/xxxxxxxxx.xxxxxxx

1 INTRODUCTION

Consider COVID-19 and the flu which share a large number of symptoms such as fever, cough, and fatigue, however, some exhibited symptoms are more prominent in either disease; e.g., anosmia is only related to COVID-19. Given that knowledge, and assuming that we have measurements of a two periods of time: one where both diseases are spreading, and one where only the flu is present, how can we identify those features/markers/symptoms that are mostly associated with the novel spreading disease COVID-19?

We focus on Google Trends data [1], and in particular the searching trends for different symptoms that are known to be common to COVID-19 and the flu, and some that are known to be mostly associated with COVID-19. This study is not the first to analyze Google Trends data, in fact, recently there has been much attention of using such data for analyzing and forecasting diseases and epidemics [2, 4, 5, 8, 10, 11]. For example, in [11], Google Trends time series are use to predict the COVID-19 cases and deaths in the United States. However, to the best of our knowledge, this is the first study that is attempting to extract knowledge about symptoms mostly related to COVID-19 by contrasting periods of time with and without the spread of COVID-19.

Given such contrasting periods of time, e.g., year 2019 and year 2020, where we monitor a set of symptoms, how can we discover which symptoms are the most *discriminative*? To that end, in this paper, we develop DNA, a novel non-negative discriminative principal component analysis, which is designed to answer the above question, and we demonstrate the viability of the method in extracting symptoms that are discriminative of COVID-19 with respect to the Flu.

2 PROBLEM FORMULATION & PROPOSED METHOD

Consider two datasets: background dataset (denoted as $\{\mathbf{y}_i \in \mathbb{R}^D\}_{i=1}^n$) which contains the information of flu, e.g., Google Trends data in 2019 or 2018 when there was no COVID-19 but flu, and target dataset (denoted as $\{\mathbf{x}_i \in \mathbb{R}^D\}_{i=1}^m$) having the information of both flu and COVID-19, e.g., Google Trends data in 2020. Here, D denotes the number of searched symptoms and i is time index. In literature, discriminative (d) principal component analysis (PCA) [3] and contrustive (c) PCA [1] performing such discriminative analysis on both the target and background datasets. Discriminative PCA seeks a projection matrix so that the ratio of the projected target data variance over that of the background data is maximized; while cPCA maximizes the difference between the target data variance and the background data variance.

Specifically, dPCA approach searches for subspace vectors, namely the columns of $\mathbf{U} \in \mathbb{R}^{D\times d}$ with $d \leq D$ by solving [3]

$$\max_{\mathbf{U}} \operatorname{Tr}\left[(\mathbf{U}^\top \mathbf{C}_y \mathbf{U})^{-1}\mathbf{U}^\top \mathbf{C}_x \mathbf{U}\right] \tag{1}$$

where $\mathbf{C}_x := \frac{1}{m}\sum_{i=1}^m (\mathbf{x}_i - \boldsymbol{\mu}_x)(\mathbf{x}_i - \boldsymbol{\mu}_x)^\top \in \mathbb{R}^{D\times D}$ representing the sample covariance of the target data with $\boldsymbol{\mu}_x$ denoting the corresponding sample mean; $\mathbf{C}_y$ is the sample covariance of the background data. This is *ratio trace* maximization problem and the columns of the optimal $\mathbf{U}$ are the right eigenvectors of

epiDAMIK 2021, Aug 15, 2021, Virtual

ACM ISBN 978-1-xxxx-XXXX-X.
https://doi.org/xx.xxxx/xxxxxxxxx.xxxxxxx

Algorithm 1: DNA.

1: **Input:** Nonzero-mean target and background data $\{\mathbf{x}_i\}_{i=1}^m$ and $\{\mathbf{y}_i\}_{i=1}^n$; number of dimensions d.
2: **Construct** covariance matrices of $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_i\}$ to obtain $\mathbf{C}_x$ and $\mathbf{C}_y$.
3: **Perform** nonnegative matrix decomposition on $\mathbf{C}_y^{-1}\mathbf{C}_x$ to obtain the two factorization components $\mathbf{W}$ and $\mathbf{H}$.
4: **Output:** $\mathbf{W}$ and $\mathbf{H}$.

$\mathbf{C}_y^{-1}\mathbf{C}_x$ associated with the top-d eigenvalues [6]. The projections $\{\mathbf{U}^\top\mathbf{x}_i \in \mathbb{R}^d\}$ are the sought lower-dimensional representations of $\{\mathbf{x}_i\}$, where the (r,l)-th entry of $\mathbf{U}$ reveals the importance of the r-th feature/symptom to the l-th projected dimension.

The challenge of using dPCA directly to uncover such symptom importance (in other words uncover discriminative symptoms of COVID-19 w.r.t. flu) is the sign ambiguity. To bypass this challenge, we propose a novel nonnegative discriminative analysis, namely DNA, by performing nonnegative matrix factorization (NNMF) on $\mathbf{C}_y^{-1}\mathbf{C}_x$. Specifically, we learn two nonnegative factorization matrices $\mathbf{W} \in \mathbb{R}^{D\times d}$ and $\mathbf{H} \in \mathbb{R}^{d\times D}$ so that

$$\mathbf{C}_y^{-1}\mathbf{C}_x \approx \mathbf{WH} \tag{2}$$

One popular approach to solve (2) is to use the Kullback–Leibler (KL) divergence metric [7]. The sought $\mathbf{W}$ will be used to estimate the importance of each symptom. Our DNA for nonnegative discriminative analytics of two datasets is summarized in Alg. 1

3 EXPERIMENTAL EVALUATION

In this section, we use a subset of the COVID-19 Search Trends symptoms dataset [9] to test the effectiveness of our proposed method. We select the number of searches of three symptoms unique for COVID-19 including *ageusia*, *shortness of breath*, and *anosmia* and six symptoms which are shared by COVID-19 and flu including *vomiting*, *diarrhea*, *cough*, *fever*, *fatigue* and *headache* from all the 51 US states in years 2018, 2019, and 2020. Our objective is to find these three unique symptoms. Throughout the experiments, we set $d = 1$.

Note that there was known spread of the COVID-19 virus in 2020 but not in 2018 or early 2019. First, we set Google Trends symptom searches in 2019 as the background data $\{\mathbf{y}_i\}_{i=1}^n$ and the searches in 2020 as the target data $\{\mathbf{x}_i\}_{i=1}^m$ with $D = 9$, $m = 19,032$ and $n = 18,980$. The resulting mean and standard derivation of the symptom coefficients (a.k.a., the column values of $\mathbf{W}$) after running the proposed DNA for 200 Monte Carlos tests which are are shown in the top left panel of Fig. 1. Similarly, we set the 2018 search data as background data and 2020 searches as target data, and plot the results in the top right panel of Fig. 1. Clearly, DNA is able to discover the discriminative symptoms of COVID-19 relative to flu. Furthermore, when we set 2020 searches as the background data and 2019 or 2018 searches as the target data; see the results in the middle panels of 1, as expected, DNA doesn't fully return the unique symptoms. This is because the unique COVID-19 symptoms are not the discriminative information of 2018/2019 data relative to 2020 data. It's worth to mention that the standard derivations of the symptoms with high coefficients in Fig. 1 are high because DNA doesn't admit unique solution and the order of the top symptoms vary a lot during different experiments. As comparison, the alternative methods such as PCA and NNMF are also tested using either 2018, 2019, or 2020 dataset; see results in Fig. 2. Contrastive (c) PCA is also carried out under different backgroud-target data setups with results in Fig. 3. One can see that neithor PCA, NNMF, nor cPCA is able to having promising discriminative analysis performance. For fairness, we also test the performance of the competing methods under different d values including $d = 1, 2, 3, 4$, and 5, and observed the symptoms coefficients for each component under each d setup, which can't find the unique COVID-19 symptoms successfully.

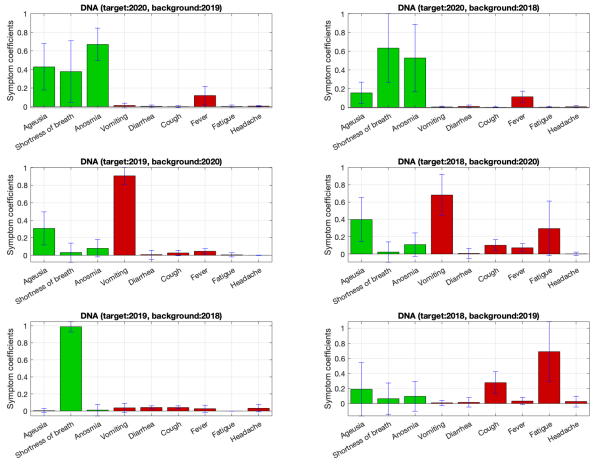

Figure 1: Symptom coefficients using DNA: DNA finds unique COVID-19 symptoms when background and target data are respectively 2018/2019 and 2020 Google Trends symptom searches.

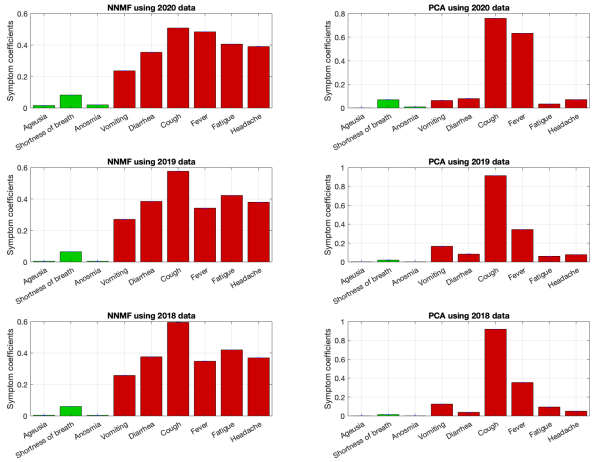

Figure 2: Symptom coefficients using NNMF and PCA which can't find unique COVID-19 symptoms.

Next, after running each method for 100 independent times while setting the background and target as the 2019 and 2020 Google

	Models	Ageusia	Shortness of breath	Anosmia	Vomiting	Diarrhea	Cough	Fever	Fatigue	Headache
Top-1 Symptom	DNA	**11**	**22**	**67**	0	0	0	0	0	0
	cPCA ($\alpha = 0.1, 0.5, 0.9$)	0	0	0	0	0	100	0	0	0
	NNMF using 2020 data	0	0	0	0	0	100	0	0	0
	NNMF using 2019 data	0	0	0	0	0	100	0	0	0
	PCA using 2020 data	0	0	0	0	0	100	0	0	0
	PCA using 2019 data	0	0	0	0	0	100	0	0	0
Top-2 Symptoms	DNA	**61**	**39**	**98**	0	0	0	2	0	0
	cPCA ($\alpha = 0.1, 0.5, 0.9$)	0	0	0	0	0	100	100	0	0
	NNMF using 2020 data	0	0	0	0	0	100	100	0	0
	NNMF using 2019 data	0	0	0	0	0	100	0	100	0
	PCA using 2020 data	0	0	0	0	0	100	100	0	0
	PCA using 2019 data	0	0	0	0	0	100	100	0	0
Top-3 Symptoms	DNA	**82**	**74**	**99**	13	1	0	19	3	9
	cPCA ($\alpha = 0.1, 0.5$)	0	0	0	0	100	100	100	0	0
	cPCA ($\alpha = 0.9$)	0	**100**	0	0	0	100	100	0	0
	NNMF using 2020 data	0	0	0	0	0	100	100	100	0
	NNMF using 2019 data	0	0	0	0	100	100	0	100	0
	PCA using 2020 data	0	0	0	0	100	100	100	0	0
	PCA using 2019 data	0	0	0	100	0	100	100	0	0

Table 1: Top-k symptom frequencies after 100 Monte Carlo experiments for different models

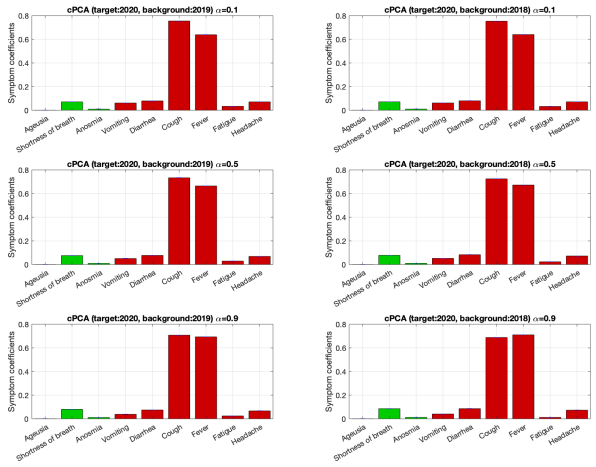

Figure 3: Symptom coefficients using cPCA which can't find unique COVID-19 symptoms.

Trends data, respectively, we investigate the frequencies of the symptoms showing up as the top-1, top-2, and top-3 by sorting the corresponding coefficients in a decreasing order. From the experiment results in Table 1, one can further conclude that the proposed DNA outperforms the existing alternatives in terms of higher frequencies of successfully searching for the discriminative symptoms.

4 CONCLUSIONS

Leveraging the advances of the discriminative principal component and the nonnegative matrix decomposition, this paper puts forward a new multiview learning model, this is terms DNA, to extract the discriminative information of one dataset relative to the other dataset. The Google COVID-19 Search Trends symptom data are used to verify the performance of the proposed method.

In the future, our research opens in several directions: (1) develop nonnegative dPCA and compare its performance against DNA; (2) understand the connection between eigenvalue decomposition and nonnegative matrix factorization on $\mathbf{C}_y^{-1}\mathbf{C}_x$; and (3) broaden the applications of DNA.

ACKNOWLEDGMENTS

Research was partially supported by the National Science Foundation Grant No. 1901379 and the Department of Defense Grant No. W911NF2110169. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding parties.

REFERENCES

[1] Abubakar Abid, Martin J Zhang, Vivek K Bagaria, and James Zou. 2017. Contrastive principal component analysis. *arXiv preprint arXiv:1709.06716* (2017).
[2] Seyed Mohammad Ayyoubzadeh, Seyed Mehdi Ayyoubzadeh, Hoda Zahedi, Mahnaz Ahmadi, and Sharareh R Niakan Kalhori. 2020. Predicting COVID-19 incidence through analysis of google trends data in iran: data mining and deep learning pilot study. *JMIR public health and surveillance* 6, 2 (2020), e18828.
[3] Jia Chen, Gang Wang, and Georgios B Giannakis. 2018. Nonlinear dimensionality reduction for discriminative analytics of multiple datasets. *IEEE Transactions on Signal Processing* 67, 3 (2018), 740–752.
[4] Maria Effenberger, Andreas Kronbichler, Jae Il Shin, Gert Mayer, Herbert Tilg, and Paul Perco. 2020. Association of the COVID-19 pandemic with internet search volumes: a Google TrendsTM analysis. *International Journal of Infectious Diseases* 95 (2020), 192–197.
[5] Atina Husnayain, Anis Fuad, and Emily Chia-Yu Su. 2020. Applications of Google Search Trends for risk communication in infectious disease management: A case study of the COVID-19 outbreak in Taiwan. *International Journal of Infectious Diseases* 95 (2020), 221–223.
[6] Ariel Jaffe and Mati Wax. 2014. Single-site localization via maximum discrimination multipath fingerprinting. *IEEE Transactions on Signal Processing* 62, 7 (2014), 1718–1728.
[7] D Lee and HS Seung. 2000. Algorithms for non˜ negative matrix factorization. In *Proc. of the 13th Int. Conf on Neural Information Processing Systems (NeurIPS 2000). New York: CurranAssociates*, Vol. 535541.
[8] Cuilian Li, Li Jia Chen, Xueyu Chen, Mingzhi Zhang, Chi Pui Pang, and Haoyu Chen. 2020. Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020. *Eurosurveillance* 25, 10 (2020), 2000199.
[9] Google LLC. 2021. Google COVID-19 Search Trends symptoms dataset. http://goo.gle/covid19symptomdataset. (Feb. 2021).
[10] Amaryllis Mavragani. 2020. Tracking COVID-19 in Europe: infodemiology approach. *JMIR public health and surveillance* 6, 2 (2020), e18941.
[11] Amaryllis Mavragani and Konstantinos Gkillas. 2020. COVID-19 predictability in the United States using Google Trends time series. *Scientific reports* 10, 1 (2020), 1–12.