Extending *t*-SNE to Obtain More Predictive Results in Healthy Singaporean Population

Suhas Vittal[1,2], Pieter Derdeyn BA[2], Kendyl Douglas BA[2], Miguel Opena[1,2], Damien Marlier MS[3], John Connolly PhD[3], David Schneider BS[2], Chris Yoo PhD[2,4]

[1]BASIS Scottsdale High School, Scottsdale, Arizona [2]Systems Imagination, Incorporated, Tempe, Arizona
[3]A*STAR, Singapore Institute for Clinical Sciences, Singapore [4]Arizona State University, Tempe, Arizona

Abstract

Visualization of high dimensional data is a challenge across all domains. With clinical trial data, patients can easily have 100s of dimensions of features, making it difficult to even look at the data, let alone gain insights from it. Machine learning techniques like t-SNE can make this easier by reducing the dimensionality of the data. Here we use an extended version of t-SNE to find and analyze clusters of patient data. We visualize the biological relationships between Singaporean young adults with respect to protein assays, in order to understand the relationship between protein expression and diabetes risk.

Introduction

One problem facing scientists today in a majority of experiments is the difficulty in identifying the characteristics of a healthy control within a large group. We use a cohort of 575 young, healthy volunteers phenotyped at the Immunomonitoring Program, A*STAR/SIgN with 116 unique serem analytes measured, detailed familial, behavioral and exposure-related questions answered [1]. Biological data visualization is a field with much promise in developing an understanding of how biological objects behave, leading to the creation of applications such as Escher, a web-based visualizer for biological pathways [2]. Similarly, nonlinear dimensionality reduction algorithms allow for the visualization of biological relationships within a dataset. Such algorithms include nonlinear Principal Components Analysis (PCA) and Autoencoder Networks [3].

In our research, we apply the dimensionality reduction algorithm *t*-Distributed Stochastic Neighbor Embedding (*t*-SNE) in order to analyze biological relationships within Singaporean young adults.

Methodology

In our data analysis, we apply the *t*-SNE algorithm, which attempts to minimize the Kullback-Leibler divergence of a lower-dimensional embedding (a compression of the original data to have less features) [4][5].

After constructing the distance matrix, we applied *t*-SNE to create a 3-dimensional embedding of the original data. We expected the "healthy" and "unhealthy" groups to segregate into two different clusters.

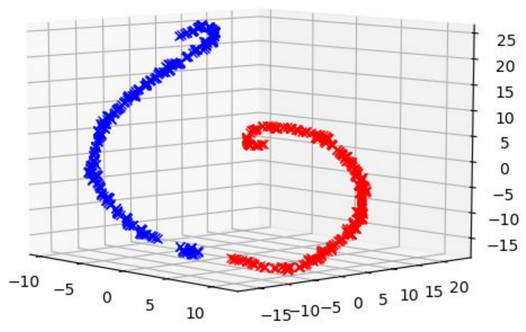

Figure 1. The 3-dimensional embedding. Blue points are healthy individuals, and red points are unhealthy individuals. The axes have no significant meaning and are purely positional.

Analysis and Results

t-SNE produced the visualization seen in Figure 1. While the healthy and unhealthy groups segregated into the blue and red curves, respectively, we uncovered an unexpected third group in between the blue and red curves which we describe as the "at risk" group. We applied the HDBSCAN clustering algorithm to identify participants in this group [6].

After identifying the three clusters, we conducted a Kruskal-Wallis test on the expression levels between the three groups to determine which varied significantly between each group. Results are in Table I.

sVEGFR3	Ab40	Ab42	ANGPTL3	ANGPTL4
ApoA1	ApoB	ApoE	ENA-78	FGF-19
FGF-21	GCP2	HCC-1	HGF	I-309
IL-16	**ApoAII**	ApoCII	ApoCIII	**Adiponectin**
Antithrombin III	BCA-1	CTACK	Eotaxin-2	Factor-XIII
I-TAC	Lymphotactin	Resistin	MIG	MIP-1d
MIP-3a	MIP-3b	NAP2	OC	OPG
PTH	SDF-1	sRAGE-2	TARC	TGFB1
Vitamin D BP	**Vitronectin**			

Table 1. 43 significant analytes from the Kruskal-Wallis test, given a level of significance of α=1e-4. Bolded analytes are prevalent in the "at risk" group.

Conclusion and Discussion

The *t*-SNE output in Figure 1 and results in Table 1 motivated the following conclusions:

1) The risk of diabetes is a continuous distribution defined as a function of the expression levels of certain proteins.
2) We discovered that 6 analytes were exclusively prevalent in the "at risk" group found in the embedding. Further, many of these analytes have recently been implicated in lipoprotein metabolism [7]. This implies that an irregularly behaving lipoprotein metabolism may be an early sign of health deterioration

For future work, we seek to get more value out of categorical data using distance hierarchies [5]. Distance hierarchies are a tool for computing more precise distances in categorical data. With a distance hierarchy, we would be able to see that people with "fish" and "shellfish" allergies are closer together than people with "fish" and "nut" allergies

We are also interested in exploring other stratifications in this dataset, including along demographics (including Chinese, Malay, and Indian ethnicities), age groups, and sex. We will use a variety of unsupervised and supervised learning techniques, including *t*-SNE, k-Nearest Neighbors, Logistic Regression, Random Forest, Decision Trees, and XG-Boost.

References

[1] e.g. in Zhisheng Her, Yiu-Wing Kam, Lisa F. P. Ng; Severity of Plasma Leakage Is Associated With High Levels of Interferon γ–Inducible Protein 10, Hepatocyte Growth Factor, Matrix Metalloproteinase 2 (MMP-2), and MMP-9 During Dengue Virus Infection,*The Journal of Infectious Diseases*, 215(1), 1 January 2017, Pages 42–51,
[2] Z. A. King, A. Dräger, A. Ebrahim, N. Sonnenschein, N. E. Lewis, and B. O. Palsson. Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways. *PLoS Comput Biol* 11(8): e1004321.
[3] A. Gisbrecht and B. Hammer. Data visualization by nonlinear dimensionality reduction. *WIREs Data Mining Knowl Discov 2015* 5, 2015:13.
[4] L. van der Maaten and G. Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 2018: 2579-2605.
[5] C-C. Hsu and W-H Huang. Integrated Dimensionality Reduction Technique for Mixed Data Involving Categorical Values. *Applied Soft Computing* 43, 2016: 199-209.
[6] R. J. G. B. Campello, D. Moulavi, A. Zimek, and J. Sander. Density-Based Clustering Based on Hierarchical Density Estimates. *Advances in Knowledge Discovery and Data Mining*. PAKDD 2013. Lecture Notes in Computer Science, vol 7819.
[7] Schwetz V, Scharnagl H, Trummer C, et al. Vitamin D supplementation and lipoprotein metabolism: A randomized controlled trial. *J Clin Lipidol*. 2018;12(3):588-596.e4.