Metric Learning for Tracking a Disease Progress

Nada Cvetkovic
Institut für Mathematik
Freie Universität Berlin
Arnimallee 6, 14195 Berlin
Email: ncvetkovic@math.fu-berlin.de

Tim Conrad
Institut für Mathematik
Freie Universität Berlin
Arnimallee 6, 14195 Berlin
Email: conrad@math.fu-berlin.de

***Abstract*—A current challenge in medicine is using high-dimensional clinical data to understand the transition of patients between long-lived stages of a disease, e.g. healthy or sick. We apply feature selection and metric learning techniques to such data in order to construct Markov chains on sparse connected networks, where long-lived stages are modeled by clusters of nodes on the network in which the chain spends long times. Using transition path theory, we can characterise the statistics of transitions and identify the most probable transition paths between long-lived stages.**

I. Introduction

The onset of a disease in a patient triggers processes that can be detected in clinical data such as blood samples. Many classification algorithms can detect the differences between healthy and diseased patients using medical samples [1]–[4]. A relevant challenge in this area is to study dynamical processes of the disease in order be able to detect intermediate states and explain the process of becoming sick. In contrast to ordinary classifiers, which can only predict the state of a patient at a single time point in the future, understanding the dynamics of the disease can predict how the patient's state will evolve over a given time interval, and thus give much more information about the patient.

In this work, we model a patient's state as a metastable Markov chain on a network. We make this choice because a patient's state depends on an extremely large number of variables whose effect can be modelled using random perturbations. Moreover, a patient's state is often assigned a class label via medical diagnoses - e.g. 'healthy' or 'sick' - where the class label evolves at a longer time scale compared to the time scale at which the patient's data evolves. For example, while the constitution of a patient's blood sample changes at the time scale of minutes, the medical diagnosis of the patient may change at the time scale of days or weeks.

II. Problem Formulation

We are given a dataset of size n consisting of pairs (x_i, y_i), $i = 1, \ldots, n$, where $x_i \in \mathbb{R}^d$ is a high-dimensional datum with $n \ll d$ and $y_i \in \{-1, +1, 0\}$ is a class label with -1, 0, and +1 corresponding to the prescribed labels 'healthy', 'transition', and 'sick' respectively. We seek to construct an irreducible Markov chain whose state space is the set $\{x_1, \ldots, x_n\}$. The problem of constructing a network from the cloud of data points is known in literature as *network inference* or *graphical model inference* [5]. It is different from the problem of network completion [6] and problem of missing link inference [7], because both these problems assume that the subset of edges is known.

The Markov chain must be metastable in the sense that, on average, it spends long times in the 'healthy' and 'sick' clusters and short times in the transition cluster. In addition, any path from the 'healthy' to the 'sick' cluster (and vice versa) contains at least one 'transition' state. Our goal is to identify the most probable transition paths from the 'healthy' to the 'sick' cluster.

To the best of our knowledge, none of the approaches for network inference aim to recover the dynamics from the static snapshot of data. On the other hand, several heuristic approaches have been developed [8]–[10] for this task. The method in [8] uses a random walk-based distance related to a diffusion map space. The methods of [9] and [10] reduce the data dimension prior to constructing the k-nearest neighbor graph, i.e. these methods use the Euclidean metric locally, but not globally, in the low dimensional space. We aim to avoid using the Euclidean metric and find a metric that better suits the geometry of the data.

III. Method

We construct a Markov chain with the desired properties by constructing a weighted, undirected, connected, sparse network from the n pairs $\{(x_i, y_i)\}_{i=1,\ldots,n}$. The problem of constructing such a network involves learning a metric that captures the geometry of the data and the information expressed by the class labels.

A. Feature Selection

Given that many metric learning methods work well on low-dimensional data, we first seek to reduce the dimension of the data vectors $\{x_i\}_{i=1,\ldots,n}$. To accomplish this, we can exploit the fact that disease biomarkers lie in a low-dimensional feature space [11]–[13].

We perform feature selection on the subset of the raw dataset $\{x_i\}_{i=1,\ldots,n}$ for which the corresponding labels are either -1 or +1 using the SPA algorithm [4]. SPA is based on the idea of one-bit compressed sensing [14]–[16] and it solves the convex program

$$\hat{\omega} = \text{argmax}_{\omega \in \mathbb{R}^d} \sum_{i=1}^{n} y_i \langle x_i, \omega \rangle \text{ subject to } ||\omega||_1 \leq \sqrt{\lambda},\ ||\omega||_2 \leq 1,$$

where λ is the sparsity controlling parameter.

The solution of the program above is a vector $\hat{\omega} \in \mathbb{R}^d$ normal to the hyperplane that best separates the subset of data with $y_i = -1$ from the subset with $y_i = +1$. Given a sufficiently small value of λ, the solution $\hat{\omega}$ is (nearly) sparse in the canonical basis. The largest (in absolute value) entries of $\hat{\omega}$ correspond to features that determine the class label y_i of each x_i with the lowest error rate. We reduce the data dimension by projecting to the dominant feature space.

B. Metric Learning and Network Construction

Once a low-dimensional feature space associated to the high-dimensional data has been identified (see Section III-A above), we perform metric learning on the low-dimensional feature space. Many approaches for metric learning have been developed, e.g. those for learning Mahalanobis distances, similarity learning, and non-linear metric learning [17]–[20]. In order to ensure that the learned metric captures the information expressed by the class labels, we use the metric learning method of Bellet et. al. [20], which can handle such global constraints. In addition, once the metric has

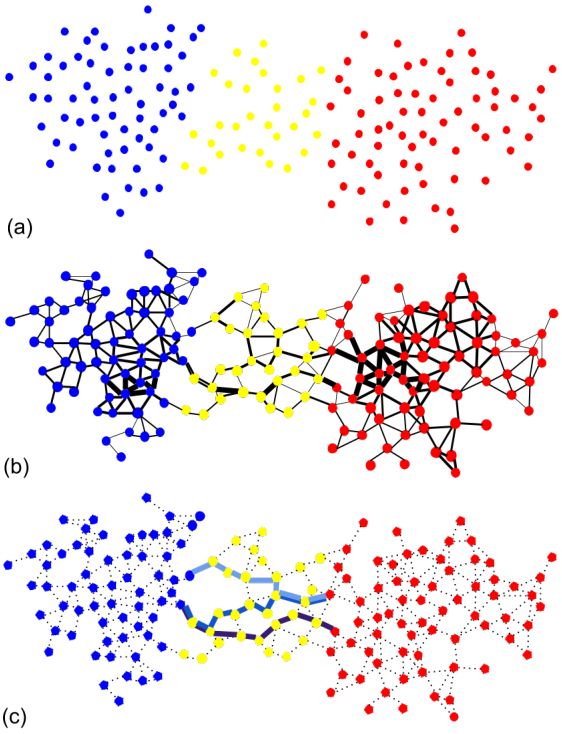

Fig. 1: A schematic that describes the approach to be taken. In subfigure (a), each labeled data point is presented as a point whose color denotes the class to which the data point belongs. Blue and red correspond to healthy and diseased clusters, respectively, and yellow corresponds to the transition region; subfigure (b), presents a constructed weighted network out of labeled samples; the three most probable paths of a disease are shown in subfigure (c) where the shade of path indicates the probability, i.e., the darker the color of the path the more probable it is.

been learned from the original data set, it can be extended and applied to additional data points. Given a learned metric, we create an undirected, weighted network, where the edge weights are computed by taking the reciprocal values of the distances between nodes(Fig. 1a,b). We apply a thresholding procedure to ensure that an edge is drawn between nodes only if they are sufficiently close with respect to the learned metric, and thus to ensure sparsity of the network.

C. Transition Path Theory

Given a Markov chain, we apply Transition Path Theory (TPT) [21], [22] in order to analyze the statistics of transition paths starting in the 'healthy' cluster and ending in the 'sick' cluster. TPT yields the probability that the Markov chain will transition according to a given transition path, and thus enables the identification of the most probable transition paths (Fig. 1c). Such qualitative analyses can then be used to propose possible treatments for a patient, for example.

Acknowledgements

N.C. and T.C. were supported by the Einstein Center for Mathematics Berlin (ECMath), project grant CH2, and by the DFG Research Center Matheon Mathematics for key technologies, Berlin. T.C. is supported by the German Ministry of Research and Education (BMBF) project Grant 3FO18501 (Forschungscampus MODAL).

References

[1] J. R. Quinlan, "Improved use of continuous attributes in c4.5," *J. Artif. Int. Res.*, vol. 4, no. 1, pp. 77–90, Mar. 1996. [Online]. Available: http://dl.acm.org/citation.cfm?id=1622737.1622742

[2] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *science*, vol. 286, no. 5439, pp. 531–537, 1999.

[3] X. Hang and F.-X. Wu, "Sparse representation for classification of tumors using gene expression data," *BioMed Research International*, vol. 2009, 2009.

[4] T. O. Conrad, M. Genzel, N. Cvetkovic, N. Wulkow, A. Leichtle, J. Vybiral, G. Kutyniok, and C. Schütte, "Sparse proteomics analysis–a compressed sensing-based approach for feature selection and classification of high-dimensional proteomics mass spectrometry data," *BMC bioinformatics*, vol. 18, no. 1, p. 160, 2017.

[5] M. Stumpf, D. J. Balding, and M. Girolami, *Handbook of statistical systems biology*. John Wiley & Sons, 2011.

[6] M. Kim and J. Leskovec, "The network completion problem: Inferring missing nodes and edges in networks," in *Proceedings of the 2011 SIAM International Conference on Data Mining*. SIAM, 2011, pp. 47–58.

[7] A. Clauset, C. Moore, and M. E. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, 2008.

[8] L. Haghverdi, F. Buettner, and F. J. Theis, "Diffusion maps for high-dimensional single-cell analysis of differentiation data," *Bioinformatics*, vol. 31, no. 18, pp. 2989–2998, 2015.

[9] S. C. Bendall, K. L. Davis, E.-a. D. Amir, M. D. Tadmor, E. F. Simonds, T. J. Chen, D. K. Shenfeld, G. P. Nolan, and D. Peer, "Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development," *Cell*, vol. 157, no. 3, pp. 714–725, 2014.

[10] J. D. Welch, A. J. Hartemink, and J. F. Prins, "Slicer: inferring branched, nonlinear cellular trajectories from single cell rna-seq data," *Genome biology*, vol. 17, no. 1, p. 106, 2016.

[11] N. Le, M. Sund, A. Vinci, G. Beyer, M. A. Javed, S. Krug, A. Neessee, and M. Schober, "Prognostic and predictive markers in pancreatic adenocarcinoma," *Digestive and Liver Disease*, vol. 48, no. 3, pp. 223 – 230, 2016.

[12] S. Tang, F. Zhou, Y. Sun, L. Wei, S. Zhu, R. Yang, Y. Huang, and J. Yang, "Cea in breast ductal secretions as a promising biomarker for the diagnosis of breast cancer: a systematic review and meta-analysis," *Breast Cancer*, pp. 1–7, 2016.

[13] A. Marrugal, L. Ojeda, L. Paz-Ares, S. Molina-Pinelo, and I. Ferrer, "Proteomic-based approaches for the study of cytokines in lung cancer," *Disease Markers*, 2016.

[14] P. T. Boufounos and R. G. Baraniuk, "1-bit compressive sensing," in *Proceedings of the 42nd Annual Conference on Information Sciences and Systems*, 2008, pp. 16–21.

[15] Y. Plan and R. Vershynin, "One-bit compressed sensing by linear programming," *Comm. Pure Appl. Math.*, vol. 66, pp. 1275–1297, 2013.

[16] Y. Plan and R. Vershynin, "Robust 1-bit compressed sensing and sparse logistic regression: a convex programming approach," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 482–494, 2013.

[17] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning with application to clustering with side-information," *Advances in neural information processing systems*, vol. 15, 2003.

[18] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, no. Feb, pp. 207–244, 2009.

[19] Y. Ying, K. Huang, and C. Campbell, "Sparse metric learning via smooth optimization," in *Advances in neural information processing systems*, 2009, pp. 2214–2222.

[20] A. Bellet, A. Habrard, and M. Sebban, "Similarity learning for provably accurate sparse linear classification," in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ser. ICML '12, J. Langford and J. Pineau, Eds. New York, NY, USA: Omnipress, July 2012, pp. 1871–1878.

[21] P. Metzner, C. Schütte, and E. Vanden-Eijnden, "Transition path theory for markov jump processes," *Multiscale Modeling & Simulation*, vol. 7, no. 3, pp. 1192–1219, 2009.

[22] N. Djurdjevac, "Methods for analyzing complex networks using random walker approaches," Ph.D. dissertation, Freie Universität Berlin Berlin, 2012.