2016 International Symposium on Nonlinear Theory and Its Applications,
NOLTA2016, Yugawara, Japan, November 27th-30th, 2016

Nonparametric clustering approach towards big data

Tom Lorimer†, Jenny Held‡, Carlo Albert‡ and Ruedi Stoop†*

†Institute of Neuroinformatics and Institute of Computational Science
University of Zurich and ETH Zurich, Winterthurerstrasse 190, 8057 Switzerland
‡Eawag, Überlandstrasse 133, 8600 Switzerland
*Email: ruedi@ini.ethz.ch

Abstract—Clustering in bioinformatics is a fundamental process involving computational issues that are far from being resolved. In our work, we propose a new approach to this problem and show preliminary comparisons to current leading methods in the field.

1. Introduction

The big data problem has infiltrated many areas of science, notably bioinformatics [1]. We focus on a specific bioinformatics problem here: the identification, discovery and interrelation of cell types. This problem has developed over recent decades into analysing automated simultaneous measurements of the abundance of tens of marker proteins on (or in) tens to hundreds of thousands of cells, most recently using mass cytometry [2]. This shift from individual to population level investigation gives rise to a new kind of difficulty in interpretation: how can structure be identified in a high dimensional space without introducing bias? It has long been known [3] that nonlinear systems give rise to convex-concave 'clusters' of similar systems (e.g. systems showing the same periodicity lie on shrimp-shaped domains in parameter space), and this has recently been suggested to manifest also in the space of observable features more generally [4]. This implies that techniques used to identify high dimensional structure in mass cytometry data need to be able to deal with convex-concave clusters. The neccessity of dealing with convex-concave clusters in mass cytometry data has also been identified recently, and a new clustering algorithm specifically proposed to deal with this problem [5]. This work will discuss our preliminary investigation of this algorithm, and compare it to our own clustering approach.

2. Toward unbiased clustering

Standard clustering approaches have a cluster shape bias that precludes accurate clustering of convex-concave sets. This bias arises from a (sometimes implicit) non-local distance criterion, where the distance from a point to a set is used to define clusters [4]. In order to cluster data without introducing bias, we need to use purely local pairwise distances between points, but still somehow 'integrate' this information to the level of a set. As a solution to this problem, Hebbian Learning Clustering (HLC) has been proposed in a previous work [4, 6]. HLC ascribes a local 'node' dynamics to each data point, and allows the dynamics of the nodes to interact via a k nearest neighbours graph. The strength of interaction across each link in the graph is weighted according to the distance between the points it connects. By exploiting a very general trade-off between the similarity of the node dynamics (homeophily), and the level of activity in the network (homeostasis), the graph's weights can self organise in an iterative manner such that the final connectivity strength of the graph determines the clusters, without requiring direct interaction across the set, and thus without introducing cluster shape bias [4, 6, 7]. HLC has recently been updated to use a more flexible and efficient map-based node dynamics defined by the Rulkov neuron model [8], and to fully exploit the sparse connectivity of the k nearest neighbour interaction matrix, rendering this approach feasible for big data problems [1]. This latest version of our algorithm, Rulkov HLC (RHLC) is used in this paper, and is described in Ref. [1].

3. Current leading approaches in mass cytometry data analysis

3.1. Visualisation: t-SNE

Student t-distributed Stochastic Neighborhood Embedding (t-SNE) [9] is a dimensionality reduction algorithm created for the visualisation of high dimensional datasets. Recently, it has been adopted in flow- and mass-cytometry data analysis under the name viSNE as an interpretation aid [10]. t-SNE achieves this dimensionality reduction by trying to match the pairwise distances between the points in the high and low dimensional spaces, where each distance is represented by a weight. Without going into detail, we note three features of this process that may cause problems for the representation of high dimensional complicated convex-concave datasets: i) the weights in the high dimensional space are normalised locally about each point, thereby removing local point density information; ii) the weight between each pair of points is made symmetric by taking the average, thereby introducing artificial inhomogeneity into the local distance information; iii) the weights in the high dimensional space are defined according to a Gaussian distribution, whereas those in the low dimensional space are defined according to a Student t-

distribution (with power law tails) resulting in a diminished sensitivity to the position of widely spaced points in the low dimensional space.

3.2. Clustering: PhenoGraph

Many clustering algorithms are currently in use on mass cytometry data (see e.g. [2, 12] for overviews). PhenoGraph [5] stands out in particular both for its claimed effectiveness and for the apparent similarity of its methodology to our own clustering algorithm that has been proposed to overcome the difficulties of standard approaches [4, 6]. PhenoGraph begins by constructing a weighted k nearest neighbours graph between the input data points. There are however two clear points of difference from HLC: i) the weights of the graph are not determined using the Euclidean distance directly, but instead using the Jaccard distance calculated on the neighbourhood overlap of the points; ii) the subdivision of the weighted graph into clusters is achieved using a well-known community detection algorithm [13]. The PhenoGraph approach has been shown to produce results that are consistent with major features identfied by manual analysis of mass cytometry data [5]. The manual analysis of mass cytometry data however, has a number of limitations. Analysis proceeds by 'manual gating': defining clusters by sequentially selecting the points within regions ('gates') in a succession of two dimensional projections of the data. Each individual gate may not necessarily be convex in the projection in which it is defined, but in all other dimensions it is convex. This places severe restrictions on the kinds of high dimensional data structures that can be identified by manual gating, and so the suitability of manual gating for high dimensional mass cytometry data has been repeatedly questioned (e.g. [5, 10]). Moreover, it is possible that either the Jaccard distance (which has a normalising effect on data density), or the objective function used in community detection algorithms (which compares weights within the entire community/cluster) might compromise the local information of the k nearest neighbours graph and lead to a cluster size or shape bias.

4. Benchmarking

We test PhenoGraph by benchmarking against custom synthetic datasets. Clustering can be seen as an 'unsupervised' task: one does not know the 'correct' answer. Therefore, before clustering algorithms can be used on real data, they need to be shown to produce stable and accurate results over a wide range of parameters on suitable test data. We base our benchmarking on synthetic two dimensional datasets, as this simplifies both the specification of arbitrarily complicated data structures, and the detailed interpretation of clustering results far beyond what is possible using standard clustering quality measures. As interesting problem settings are high dimensional, we generate high dimensional test datasets by embedding two dimensional datasets in higher dimensions. This approach permits the detailed analysis of high dimensional clustering results in the original two dimensional space. Using synthetic data for benchmarking (as opposed to, e.g. manually gated mass cytometry data) moreover guarantees the accuracy of the test labels.

To provide an overview of the clustering results across a range of parameters, we use the standard F_1 score or F-measure, i.e. the harmonic mean between 'precision' and 'recall' of a given cluster i with respect to a retreived cluster j

$$F_{ij} = 2\frac{f_p f_r}{f_p + f_r}\,, \tag{1}$$

where the precision, f_p is the fraction of points in the retrieved cluster j that are correctly assigned to given cluster i, and the recall, f_r is the fraction of the points in the given cluster i that are assigned to the retrieved cluster j. For each given cluster i, F_{ij} will be different for different j. We define $F_i = \max_j F_{ij}$, and as an overall characterisation of the clustering, take either a mean, giving the unweighted F-measure $F = \frac{1}{n}\sum_i F_i$, or a weighted mean, giving the weighted F-measure $F_w = \sum_i \frac{|i|}{N} F_i$, where n is the number of given clusters i, and N is the total number of data points. These are standard statistical measures used for the assessment of clustering algorithms in general, including cytometry clustering algorithms [11, 12].

4.1. Two dimensions

We generated a suite of datasets of varying difficulty, each containing convex-concave shapes, with varying degrees of background noise. Selected here as an illustrative (rather than representative) example is a dataset of two concentric rings, with equal uniform density, separated by a thin band of lower density uniform noise (when calculating the F-measure, the assignment of points in the band of low density noise was ignored). We see in Fig. 1 that PhenoGraph does not successfully cluster this dataset for any tested value of k. Despite claims to the contrary, there is a clear cluster shape/size bias that precludes the inclusion of the entire outer ring in one cluster before the inner ring is also included. RHLC, by contrast, can successfully deal with this problem for a wide range of parameters (Fig. 2).

4.2. Higher dimensions

Our high dimensional test dataset with convex-concave structures for benchmarking, is composed from test datasets in 2 dimensions of differing sizes and densities highlighting a range of different difficulties that may be faced in clustering natural data. We transformed this two dimensional composed dataset into 8 dimensions according to

$$(x, y) \to (x + y, x - y, x^2, y^2, xy, x^2 y, xy^2, x^3 y^2)\,, \tag{2}$$

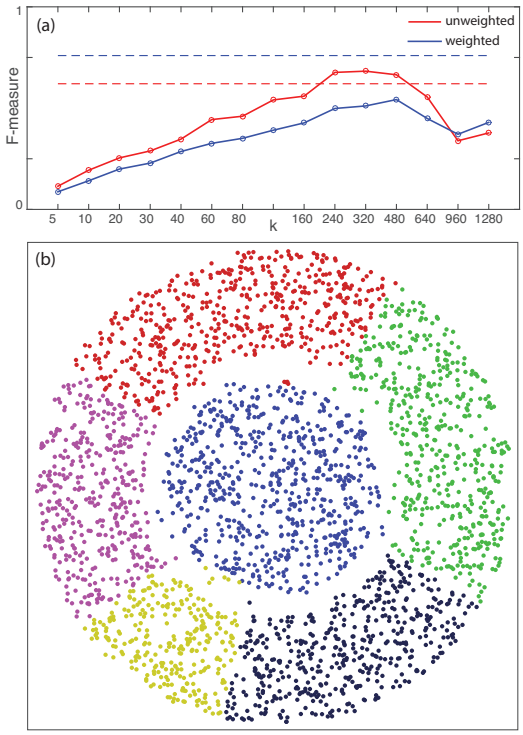

Figure 1: PhenoGraph performance on two dimensional dataset. a) F-measure as a function of the only algorithm paramter, k, dashed lines indicate F-measure for case where all points belong to the same cluster. b) Example clustering result for $k = 320$, retrieved clusters indicated by colours.

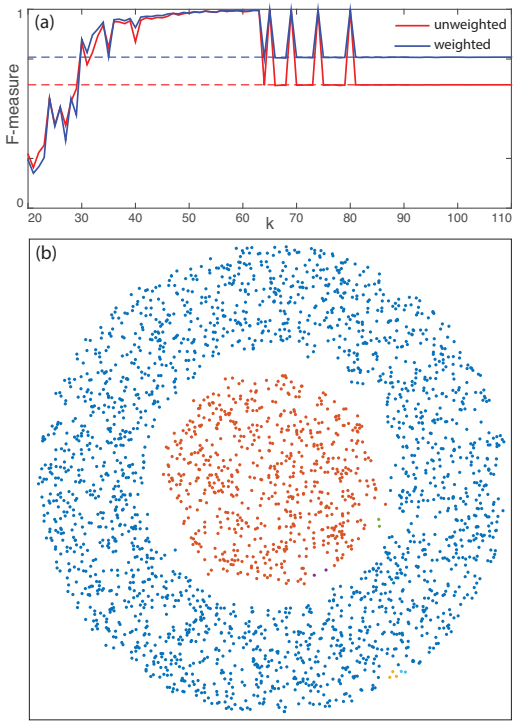

Figure 2: RHLC performance on two dimensional dataset. a) F-measure as a function of primary algorithm parameter: number of nearest neighbours k. b) Example clustering result for $k = 63$, retrieved clusters indicated by colours.

such that the original 2 dimensional dataset now sits on a 2 dimensional sub-manifold of an 8 dimensional space. Although the first two dimensions of the transformation simply apply a rotation to the original dataset (so that there exists a projection that retains the original structure), this a priori knowledge is not available to the algorithms we test.

To illustrate the difficulty of reverse transforming such convex-concave data from a high dimensional space to two dimensions, even in the case where they are known to lie on a 2 dimensional sub-manifold, we performed a t-SNE transformation of our high dimensional test data set [9]. While it is not to be expected that t-SNE should return the original 2 dimensional configuration of points, we found that the t-SNE transformed data could not reasonably be interpreted in a way that would return the correct point labeling. While the major convex sets were preserved, the major convex-concave sets were partitioned in such a way that the pieces were no longer adjacent in the two dimensional space. More complicated convex-concave sets were partitioned into many pieces spread across the two dimensional plane, illustrating the difficulty of using t-SNE tranformed data for interpretation.

Testing PhenoGraph on our high dimensional dataset, we found that it suffered similar problems to the two dimensional case, namely, an inherent cluster size/shape bias as a function of its parameter (Fig. 3). Although the weighted F-measure appears to monotonically increase across the tested range, we observe that this already coincides with an incorrect coarse grouping and splitting of clusters that can be expected to deteriorate further with further increasing k.

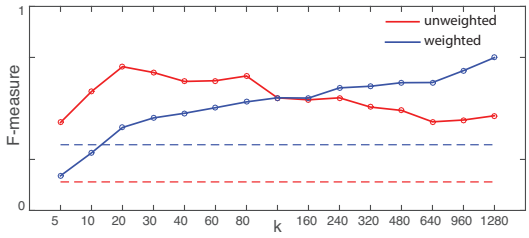

Figure 3: PhenoGraph performance on high dimensional dataset. F-measure as a function of the only algorithm paramter, k, dashed lines indicate F-measure for the case where all points belong to the same cluster.

RHLC avoids this inherent cluster size/shape bias, and can successfully cluster the data over a wide range of parameters (Fig. 4). Even for RHLC however, this dataset is exceptionally difficult. RHLC has no local density normalisation, and we note that it struggles to cluster the lowest density cluster. This points the way toward a sequential clustering approach in future implementations.

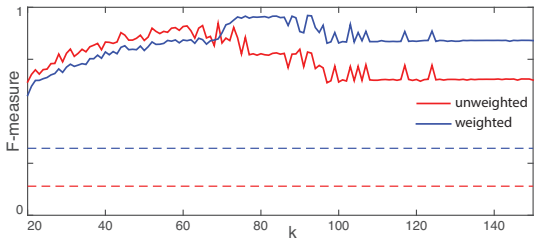

Figure 4: RHLC performance on high dimensional dataset. F-measure as a function of primary algorithm parameter: number of nearest neighbours k, dashed lines indicate F-measure for case where all points belong to the same cluster.

5. Real data and outlook

The synthetic data examples presented so far provide important insight into the limitations of PhenoGraph and t-SNE, but how do these manifest when they are applied to real data? We are currently testing RHLC on one of the datasets used to benchmark PhenoGraph: a mass cytometry dataset of healthy human bone marrow cells described in Ref. [14]. We find that RHLC consistently merges some large clusters that were split both by manual gating and PhenoGraph. However based on our synthetic results, where PhenoGraph made artificial partitions of the clusters, we are currently investigating whether this is an RHLC clustering error, or whether these groups of cell types are actually joined in this dataset in a continuum of cell differentiation in high dimensions.

Acknowledgments

This work was supported by SNF grant numbers CR32I3 159660 to C.A. and 200021-153542/1 to R.S..

References

[1] R. Stoop, K. Kanders, T. Lorimer, J. Held and C. Albert, "Big data naturally rescaled," *Chaos Soliton. Fract.*, doi:10.1016/j.chaos.2016.02.035, 2016.

[2] F. Mair et al., "The end of gating? An introduction to automated analysis of high dimensional cytometry data," *Eur. J. Immunol.*, vol.46, pp.34–43, 2016.

[3] R. Stoop, P. Benner and Y. Uwate, "Real-world existence and origins of global shrimp organization on spirals," *Phys. Rev. Lett.*, vol.105, p.074102, 2010.

[4] F. Gomez, R.-L. Stoop and R. Stoop, "Universal dynamical features preclude standard clustering in a large class of biochemical data," *Bioinformatics*, vol.30, pp.2486–2493, 2014.

[5] J. H. Levine et al., "Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis," *Cell*, vol.162, pp.184–197, 2015.

[6] F. Landis, T. Ott and R. Stoop, "Hebbian self-organizing integrate-and-fire networks for data clustering," *Neural Computation*, vol.22, pp.273–288, 2010.

[7] R. Gutiérrez, et al., "Emerging meso- and macroscales from synchronization of adaptive networks," *Phys. Rev. Lett.*, vol.107, p.234103, 2011.

[8] N. F. Rulkov, "Modeling of spiking-bursting neural behavior using two-dimensional map," *Phys. Rev. E*, vol.65, p.041922, 2002.

[9] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol.9, pp.2579–2605, 2008.

[10] E.-A. D. Amir et al., "viSNE enables visualization of high demensional single-cell data and reveals phenotypic heterogeneity of leukemia," *Nat. Biotechnol.*, vol.31, pp.545–552, 2013.

[11] N. Aghaeepour et al., "Critical assessment of automated flow cytometry data analysis techniques," *Nat. Methods*, vol.10, pp.228–138, 2013.

[12] L. M. Weber and M. D. Robinson, "Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data," *bioRxiv*, doi:10.1101/047613, 2016.

[13] V. D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech.*, p.P10008, 2008.

[14] S. C. Bendall et al., "Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum," *Science* vol.332, pp.687–696, 2011.