Branch points and stability

J.F. Jardine*

May 3, 2020

Abstract

The hierarchy and branch point posets for a data set each have a calculus of least upper bounds. Upper bounds are used to show that the map of branch points associated to the inclusion of data sets is a controlled homotopy equivalence, where the control is defined by an upper bound relation that is associated to their Hausdorff distance.

Introduction

This paper is a discussion of clustering phenomena that arise in connection with inclusions $X \subset Y \subset \mathbb{R}^n$ of data sets, interpreted through the lens of hierarchies of clusters and branch points.

Suppose that X is a finite subset (a data set) in a metric space Z. There is a well known system of simplicial complexes $V_s(X)$ whose simplices are the subsets σ of X such that $d(x, y) \leq s$ for each pair of points $x, y \in \sigma$, where d is the metric on Z. The complexes $V_s(X)$ are the Vietoris-Rips complexes for the data set X.

If k is a positive integer, $L_{s,k}(X)$ is the subcomplex of $V_s(X)$ whose simplices σ have vertices x such that $d(x, y) \leq s$ for at least k distinct points $y \neq x$ in X. This object is variously called a degree Rips complex, or a Lesnick complex. The number k is a density parameter.

The simplicial complexes $V_s(X)$ and $L_{s,k}(X)$ are defined by their respective partially ordered sets (posets) of simplices $P_s(X)$ and $P_{s,k}(X)$ [4]. The corresponding nerves $BP_s(X)$ and $BP_{s,k}(X)$ are barycentric subdivisions of the respective complexes $V_s(X)$ and $L_{s,k}(X)$, and therefore have the same homotopy types. This identification of homotopy types is assumed in this paper, so that $V_s(X) = BP_s(X)$ and $L_{s,k}(X) = BP_{s,k}(X)$, respectively.

A relationship $s \leq t$ between spatial parameters induces an inclusion

$$L_{s,k}(X) \subset L_{t,k}(X).$$

2020 Mathematics Subject Classification:: Primary 55U10; Secondary 62R40, 68T09

*Supported by NSERC.

Some of the complexes $L_{s,k}(X)$ could be empty, and $L_{s,k}(X)$ is the barycentric subdivision of a big simplex for s sufficiently large if k is bounded above by the the cardinality of X. Observe also that $L_{s,0}(X) = V_s(X)$, and that the subobjects $L_{s,k}(X)$ filter $V_s(X)$.

For a fixed integer k, the sets $\pi_0 L_{s,k}(X)$ of path components, as s varies, define a tree $\Gamma_k(X)$ with elements $(s, [x])$ such that $[x] \in \pi_0 L_{s,k}(X)$.

The tree $\Gamma_k(X)$ is the object studied by the HDBSCAN clustering algorithm, while the individual sets of clusters $\pi_0 L_{s,k}(X)$ are computed for the DBSCAN algorithm.

The tree $\Gamma_k(X)$ has a subobject $\mathrm{Br}_k(X)$ whose elements are the branch points of the tree $\Gamma_k(X)$. The branch points of $\Gamma_k(X)$ are in one to one correspondence with the stable components for $\Gamma_k(X)$ that are defined in [3], in the sense that every stable component starts at a unique branch point. We replace the stable component discussion of [3] with the branch point tree $\mathrm{Br}_k(X)$, and make particular use of its ordering.

The branch point tree $\mathrm{Br}_k(X)$ is a highly compressed version of the hierarchy $\Gamma_k(X)$ that is produced by the HDBSCAN algorithm.

We derive a stability result (Theorem 2) for the branch point tree. This result follows from a stability theorem for the degree Rips complex [4], together with a calculus of least upper bounds for the branch point tree that is developed in the next section.

Suppose that $i : X \subset Y$ are data sets in a metric space Z, and that $r > 0$. Suppose that the Hausdorff distance $d_H(X^{k+1}_{dis}, Y^{k+1}_{dis}) < r$ in Z^{k+1}, where X^{k+1}_{dis} is the set of $k+1$ distinct points in X, interpreted as a subset of the product metric space Z^{k+1}. The inclusion i induces an inclusion $i : L_{s,k}(X) \to L_{s,k}(Y)$ of simplicial complexes, which is natural in all s and k.

The stability theorem for the degree Rips complex (Theorem 6 of [4], which is a statement about posets) implies the following:

Theorem 1. *Suppose that $X \subset Y \subset Z$ are data sets, and we have the relation*

$$d_H(X^{k+1}_{dis}, Y^{k+1}_{dis}) < r$$

on Hausdorff distance between associated configuration spaces in Z^{k+1}. Then there is a diagram of simplicial complex maps

$$\begin{array}{ccc} L_{s,k}(X) & \xrightarrow{\sigma} & L_{s+2r,k}(X) \\ {\scriptstyle i}\downarrow & \overset{\theta}{\nearrow} & \downarrow{\scriptstyle i} \\ L_{s,k}(Y) & \xrightarrow[\sigma]{} & L_{s+2r,k}(Y) \end{array} \qquad (1)$$

in which the horizontal and vertical maps are natural inclusions. The upper triangle of the diagram commutes, and the lower triangle commutes up to a homotopy which fixes $L_{s,k}(X)$.

Theorem 1 specializes to the Rips stability theorem in the case $k = 0$ (see [4], [1]). The picture (1) is often called a homotopy interleaving.

Application of the path component functor π_0 to the diagram (1) gives a commutative diagram

$$\begin{array}{ccc} \pi_0 L_{s,k}(X) & \xrightarrow{\sigma} & \pi_0 L_{s+2r}(X) \\ {\scriptstyle i}\downarrow & \nearrow^{\theta} & \downarrow{\scriptstyle i} \\ \pi_0 L_{s,k}(Y) & \xrightarrow[\sigma]{} & \pi_0 L_{s+2r}(Y) \end{array} \tag{2}$$

which is an interleaving of clusters. This is true for all homotopy invariants: in particular, application of homology functors to (1) produces interleaving diagram in homology groups.

The tree $\Gamma_k(X)$ has least upper bounds, and these restrict to least upper bounds for the subtree $\mathrm{Br}_k(X)$ of branch points (Lemma 3).

The inclusion $\mathrm{Br}_k(X) \subset \Gamma_k(X)$ is a homotopy equivalence of posets, where the homotopy inverse is defined by taking the maximal branch point $(s_0, [x_0]) \leq (s, [x])$ below $(s, [x])$ for each object of $\Gamma_k(X)$. The existence of the maximal branch point below an object $(s, [x])$ is a consequence of Lemma 6.

The poset map $i : \Gamma_k(X) \to \Gamma_k(Y)$ defines a poset map $i_* : \mathrm{Br}_k(X) \to \mathrm{Br}_k(Y)$, via the homotopy equivalences for the data sets X and Y of the last paragraph. The maps $\theta : \pi_0 L_{s,k}(Y) \to \pi_0 L_{s+2r}(X)$ similarly induce morphisms of trees $\theta_* : \Gamma_k(Y) \to \Gamma_k(X)$ and $\theta_* : \mathrm{Br}_k(Y) \to \mathrm{Br}_k(X)$.

We then have the following:

Theorem 2. *Under the assumptions of Theorem 1, there is a homotopy commutative diagram*

$$\begin{array}{ccc} \mathrm{Br}_k(X) & \xrightarrow{\sigma_*} & Br_k(X) \\ {\scriptstyle i_*}\downarrow & \nearrow^{\theta_*} & \downarrow{\scriptstyle i_*} \\ \mathrm{Br}_k(Y) & \xrightarrow[\sigma_*]{} & \mathrm{Br}_k(Y) \end{array} \tag{3}$$

of morphisms of trees.

Theorem 2 is a stability theorem for branch points, when interpreted in the language that is developed here.

In particular, the homotopies of Theorem 2 are given by relations

$$\theta_* i_*(s, [x]) \leq \sigma_*(s, [x])$$

and

$$i_* \theta_*(t, [y]) \leq \sigma_*(t, [y])$$

for branch points $(s, [x]) \in \mathrm{Br}_k(X)$ and $(t, [y]) \in \mathrm{Br}_k(Y)$, respectively.

Then, for example, $\sigma_*(x, [x])$ is the maximal branch point below $(s+2r, [x])$, so that $(s, [x]) \leq \sigma_*(s, [x]) \leq (s + 2r, [x])$. It follows that the branch points

$(s,[x])$ and $\theta_* i_*(s,[x])$ have a least upper bound that is lies between $(s,[x])$ and $(s+2r,[x])$. This is a significant constraint on the placement of that upper bound if r is small.

A similar constraint exists for the least upper bound of the points $(t,[y])$ and $i_*\theta_*(t,[y])$ in $\mathrm{Br}_k(Y)$.

1 Branch points and upper bounds

Fix the density number k and suppose that $L_{s,k}(X) \neq \emptyset$ for s sufficiently large. Apply the path component functor to the $L_{s,k}(X)$, to get a diagram of functions

$$\dots \to \pi_0 L_{s,k}(X) \to \pi_0 L_{t,k}(X) \to \dots$$

The graph $\Gamma_k(X)$ has vertices $(s,[x])$ with $[x] \in \pi_0 L_{s,k}(X)$, and edges $(s,[x]) \to (t,[x])$ with $s \leq t$. This graph underlies a poset with a terminal object, and is therefore a tree (or hierarchy).

The morphisms of $\Gamma_k(X)$ are relations $(s,[x]) \leq (t,[y])$. The existence of such a relation means that $[x] = [y] \in \pi_0 L_{t,k}(X)$, or that the image of $[x] \in \pi_0 L_{s,k}(X)$ is $[y]$ under the induced function $\pi_0 L_{s,k}(X) \to \pi_0 L_{t,k}(X)$.

Remarks: 1) Partitions of X given by the set $\pi_0 V_s(X)$ are standard clusters. The tree $\Gamma_0(X) = \Gamma(V_*(X))$ defines a hierarchical clustering that is similar to the single linkage clustering.

2) The set $\pi_0 L_{s,k}(X)$ gives a partitioning of the set of elements of X having at least k neighbours of distance $\leq s$, which is the subject of the DBSCAN algorithm. The tree $\Gamma_k(X) = \Gamma(\pi_0 L_{*,k}(X))$ is the structural object underlying the HDBSCAN algorithm.

A *branch point* in the tree $\Gamma_k(X)$ is a vertex $(t,[x])$ such that either of following two conditions hold:

1) there is an $s_0 < t$ such that for all $s_0 \leq s < t$ there are distinct vertices $(s,[x_0])$ and $(s,[x_1])$ with $(s,[x_0]) \leq (t,[x])$ and $(s,[x_1]) \leq (t,[x])$, or

2) there is no relation $(s,[y]) \leq (t,[x])$ with $s < t$.

The second condition means that a representing vertex x of the path component $[x] \in \pi_0 L_{t,k}(X)$ is not a vertex of $L_{s,k}(X)$ for $s < t$. Write $\mathrm{Br}_k(X)$ for the set of branch points $(s,[x])$ in $\Gamma_k(X)$.

The set $\mathrm{Br}_k(X)$ inherits a partial ordering from the poset $\Gamma_k(X)$, and the inclusion $\mathrm{Br}_k(X) \subset \Gamma_k(X)$ of the set of branch points defines a monomorphism of posets.

Every branch point $(s,[x])$ of $\Gamma_k(X)$ has $s = s_i$, where s_i is a phase change number for X. The phase change numbers are the various distances $d(x,y)$ between the elements of the finite set X.

The branch point poset $\mathrm{Br}_k(X)$ is a tree, because the element $(s,[x])$ corresponding to the largest phase change number s is terminal.

Suppose that $(s,[x])$ and $(t,[y])$ are vertices of the graph $\Gamma_k(X)$. There is a vertex $(v,[w])$ such that $(s,[x]) \leq (v,[w])$ and $(t,[y]) \leq (v,[w])$. The two relations specify that $[x]=[z]=[y]$ in $\pi_0 L_{v,k}(X)$.

There is a unique smallest vertex $(u,[z])$ which is an upper bound for both $(s,[x])$ and $(t,[y])$. The number u is the smallest parameter (necessarily a phase change number) such that $[x]=[y]$ in $\pi_0 L_{u,k}(X)$, and so $[z]=[x]=[y]$. In this case, one writes

$$(s,[x]) \cup (t,[y]) = (u,[z]).$$

The vertex $(u,[z])$ is the *least upper bound* (or join) of $(s,[x])$ and $(t,[y])$.

Every finite collection of points $(s_1,[x_1]),\dots,(s_p,[x_p])$ has a least upper bound

$$(s_1,[x_1]) \cup \dots \cup (s_p,[x_p])$$

in the tree $\Gamma_k(X)$.

Lemma 3. *The least upper bound $(u,[z])$ of branch points $(s,[x])$ and $(t,[y])$ is a branch point.*

Proof. If there is a number v such that $s,t < v < u$, then $(v,[x])$ and $(v,[y])$ are distinct because $(u,[z])$ is a least upper bound, so that $(u,[z])$ is a branch point.

Otherwise, $s=u$ or $t=u$, in which case $(u,[z])=(s,[x])$ or $(u,[z])=(t,[y])$. In either case, $(u,[z])$ is a branch point. □

It follows from Lemma 3 that any two branch points $(s,[x])$ and $(t,[y])$ have a least upper bound in $\mathrm{Br}_k(X)$, and that the poset inclusion $\alpha : \mathrm{Br}_k(X) \to \Gamma_k(X)$ preserves least upper bounds.

We have the following observation:

Lemma 4. *Suppose that $(s_1,[x_1]),(s_2,[x_2])$ and $(s_3,[x_3])$ are vertices of $\Gamma_k(X)$. Then*

$$(s_1,[x_1]) \cup (s_3,[x_3]) \leq ((s_1,[x_1]) \cup (s_2,[x_2])) \cup ((s_2,[x_2]) \cup (s_3,[x_3])).$$

Remark: Carlsson and Mémoli [2] define an ultrametric d on $X=V_0(X)$, for which they say that $d(x,y)=s$, where s is the minimum parameter value such that $[x]=[y] \in \pi_0 V_s(X)$.

The least upper bound concept is both an extension of and a potential replacement for this ultrametric, and Lemma 4 is the analog for the triangle inequality.

The Carlsson-Mémoli theory does not apply to the full tree $\Gamma_k(X)$, because the vertex sets of the Lesnick complexes $L_{s,k}(X)$ can vary with changes of the distance parameter s. We can, however, define an ultrametric on each of the sets $\pi_0 L_{s,k}(X)$ as follows:

Suppose given $[x]$ and $[y]$ in $\pi_0 L_{s,k}(X)$ (or equivalently, points $(s,[x])$ and $(s,[y])$ in $\Gamma_k(X)$). Write $d([x],[y]) = u - s$, where $(s,[x]) \cup (s,[y]) = (u,[w])$.

Lemma 5. *Every vertex $(s, [x])$ of $\Gamma_k(X)$ has a unique largest branch point $(s_0, [x_0])$ such that $(s_0, [x_0]) \leq (s, [x])$.*

Proof. The least upper bound of the finite list of the branch points $(t, [y])$ such that $(t, [y]) \leq (s, [x])$ is a branch point, by Lemma 3. □

In the situation of Lemma 5, one says that $(s_0, [x_0])$ is the *maximal branch point below* $(s, [x])$.

If $(s, [x])$ is a branch point, then the maximal branch point below $(s, [x])$ is $(s, [x])$, by construction.

Lemma 6. *Suppose that $(s_0, [x_0])$ and $(t_0, [y_0])$ are maximal branch points below the points $(s, [x])$ and $(t, [y])$ in $\Gamma_k(X)$, respectively. Then $(s_0, [x_0]) \cup (t_0, [y_0])$ is the maximal branch point below $(s, [x]) \cup (t, [y])$.*

Proof. Suppose that $s \leq t$.

We have

$$(s_0, [x_0]) \cup (t_0, [y_0]) \leq (s, [x]) \cup (t, [y]).$$

and $(s_0, [x_0]) \cup (t_0, [y_0])$ is a branch point by Lemma 3.

Write

$$(v, [z]) = (s_0, [x_0]) \cup (t_0, [y_0]).$$

1) Suppose that $v \leq t$. Then

$$(t_0, [y_0]) \leq (t, [y]) = (t, [y_0])$$

and

$$(t_0, [y_0]) \leq (v, [z]) = (v, [y_0]),$$

so that

$$(v, [z]) = (v, [y_0]) \leq (t, [y_0]) = (t, [y])$$

since $v \leq t$.

Also, $(s_0, [x_0]) \leq (s, [x])$ and $(s_0, [x_0]) \leq (v, [z]) \leq (t, [y])$ so that $(s, [x]) \leq (t, [y])$.

Then $(s_0, [x_0]) \leq (t_0, [y_0])$ by maximality, and it follows that

$$(s_0, [x_0]) \cup (t_0, [y_0]) = (t_0, [y_0])$$

is the maximal branch point below

$$(s, [x]) \cup (t, [y]) = (t, [y])$$

2) Suppose that $v > t$. Then $(s, [x]) = (s, [x_0]) \leq (v, [z])$ and $(t, [y]) = (t, [y_0]) \leq (v, [z])$ because $s \leq t < v$, so that

$$(s, [x]) \cup (t, [y]) \leq (s_0, [x_0]) \cup (t_0, [y_0]),$$

Thus, $(s_0, [x_0]) \cup (t_0, [y_0]) = (s, [x]) \cup (t, [y])$ is a branch point, by Lemma 3. □

Lemma 7. *The poset inclusion* $\alpha : \mathrm{Br}_k(X) \to \Gamma_k(X)$ *has an inverse*

$$max : \Gamma_k(X) \to \mathrm{Br}_k(X),$$

up to homotopy, and $\mathrm{Br}_k(X)$ *is a strong deformation retract of* $\Gamma_k(X)$.

Proof. Lemma 5 implies that every vertex $(s, [x])$ of $\Gamma_k(X)$ has a unique maximal branch point $(s_0, [x_0])$ such that $(s_0, [x_0]) \leq (s, [x])$. Set

$$max(s, [x]) = (s_0, [x_0]).$$

The maximality condition implies that max preserves the ordering. The composite $max \cdot \alpha$ is the identity on $\mathrm{Br}_k(X)$, and the relations $(s_0, [x_0]) \leq (s, x)$ define a homotopy $max \cdot \alpha \leq 1$ that restricts to the identity on $\mathrm{Br}_k(X)$. □

Return to the inclusion $i : X \subset Y \subset \mathbb{R}^n$ of finite data sets. Suppose that $d_H(X^{k+1}_{dis}, Y^{k+1}_{dis}) < r$ and that $L_{s,k}(Y)$ is non-empty, as in the statement of Theorem 1.

Write $i_* : \mathrm{Br}_k(X) \to \mathrm{Br}_k(Y)$ for the composite poset morphism

$$\mathrm{Br}_k(X) \xrightarrow{\alpha} \Gamma_k(X) \xrightarrow{i_*} \Gamma_k(Y) \xrightarrow{max} \mathrm{Br}_k(Y)$$

This map takes a branch point $(s, [x])$ to the maximal branch point below $(s, [i(x)])$.

Remark: The map $i_* : \mathrm{Br}_k(X) \to \mathrm{Br}_k(Y)$ only preserves least upper bounds up to homotopy. Suppose that $(s, [x])$ and $(t, [y])$ are branch points of X, and let $(s_0, [x_0]) \leq (s, [i(x)])$ and $(t_0, [y_0]) \leq (t, [i(y)])$ be maximal branch points below the images of $(s, [x])$ and $(t, [y])$ in $\Gamma_k(Y)$. Then $(s_0, [x_0]) \cup (t_0, [y_0])$ is the maximal branch point below $(s, [i(x)]) \cup (t, [i(y)])$ by Lemma 6, but it may not be the maximal branch point below $i_*((s, [x]) \cup (t, [y]))$.

Poset morphisms $\theta_* : \mathrm{Br}_k(Y) \to \mathrm{Br}_k(X)$ and $\sigma_* : \mathrm{Br}_k(X) \to \mathrm{Br}_k(X)$ are similarly defined, by the poset morphism $\theta : \Gamma_k(Y) \to \Gamma_k(X)$ given by $(t, [y]) \mapsto (t + 2r, [\theta(y)])$, and the shift morphism $\sigma : \Gamma_k(X) \to \Gamma_k(X)$ given by $(s, [x]) \mapsto (s + 2r, [x])$. These maps again preserve least upper bounds up to homotopy.

1) Consider the poset maps

$$\mathrm{Br}_k(X) \xrightarrow{i_*} \mathrm{Br}_k(Y) \xrightarrow{\theta_*} \mathrm{Br}_k(X).$$

If $(s, [x])$ is a branch point for X, choose maximal branch points $(s_0, [x_0]) \leq (s, [i(x)]$ for Y, $(s_1, [x_1]) \leq (s_0 + 2r, [\theta(x_0)])$ and $(v, [y]) \leq (s + 2r, [x])$ below the respective objects.

Then $\theta_* i_*(s, [x]) = (s_1, [x_1])$, and there is a natural relation

$$\theta_* i_*(s, [x]) = (s_1, [x_1]) \leq (v, [y]) = \sigma_*(s, [x])$$

by a maximality argument. We therefore have a homotopy of poset maps

$$\theta_* i_* \leq \sigma_* : \mathrm{Br}_k(X) \to \mathrm{Br}_k(X). \tag{4}$$

2) Similarly, if $(t,[y])$ is a branch point of Y, then

$$i_*\theta_*(t,[y]) \leq \sigma_*(t,[y]),$$

giving a homotopy

$$i_*\theta_* \leq \sigma_* : \mathrm{Br}_k(Y) \to \mathrm{Br}_k(Y). \tag{5}$$

The construction of the poset maps i_*, θ_* and σ_*, together with the relations (4) and (5), complete the proof of Theorem 2.

There are relations

$$(s,[x]) \leq \sigma_*(s,[x]) \leq (s+2r,[x]) \tag{6}$$

for branch points $(s,[x])$. It follows that the poset map $\sigma_* : \mathrm{Br}_k(X) \to \mathrm{Br}_k(X)$ is homotopic to the identity on $\mathrm{Br}_k(X)$.

It also follows that $\sigma_*(s,[x]) = (t,[x])$ is close to $(s,[x])$ in the sense that $t - s \leq 2r$. Thus, the branch points $(s,[x])$ and $\theta_* i_*(s,[x])$ have a common upper bound, namely $\sigma_*(s,[x])$, which is close to $(s,[x])$.

The subobject of $\mathrm{Br}_k(X)$ consisting of all branch points of the form $(s,[x])$ as s varies has an obvious notion of distance: the distance between points $(s,[x])$ and $(t,[x])$ is $|t-s|$.

If $(t,[y])$ is a branch point of $\Gamma_k(Y)$, the branch point $\sigma_*(t,[y])$ is similarly an upper bound for $(t,[y])$ and $i_*\theta_*(t,[y])$ that is close to $(t,[y])$.

References

[1] Andrew J. Blumberg and Michael Lesnick. Universality of the homotopy interleaving distance. *CoRR*, abs/1705.01690, 2017.

[2] Gunnar Carlsson and Facundo Mémoli. Characterization, stability and convergence of hierarchical clustering methods. *J. Mach. Learn. Res.*, 11:1425–1470, 2010.

[3] J.F. Jardine. Stable components and layers. Canad. Math. Bull., doi:10.4153/S000843951900064X, 2019.

[4] J.F. Jardine. Persistent homotopy theory. Preprint, arxiv: 2002:10013 [math.AT], 2020.