PERSISTENT TOPOLOGY OF DECISION BOUNDARIES

Kush R. Varshney and Karthikeyan Natesan Ramamurthy

Mathematical Sciences and Analytics Department, IBM Thomas J. Watson Research Center
1101 Kitchawan Road, Yorktown Heights, NY 10598, USA

ABSTRACT

Topological signal processing, especially persistent homology, is a growing field of study for analyzing sets of data points that has been heretofore applied to unlabeled data. In this work, we consider the case of labeled data and examine the topology of the decision boundary separating different labeled classes. Specifically, we propose a novel approach to construct simplicial complexes of decision boundaries, which can be used to understand their topology. Furthermore, we illustrate one use case for this line of theoretical work in kernel selection for supervised classification problems.

***Index Terms*—** Graph walk, persistent homology, simplicial complex, supervised classification, topological data analysis

1. INTRODUCTION

Topology is the mathematical study of shape. Topological signal processing is finding application in many different domains and problems, including multiple target detection and localization, tracking in water pollution analysis, testing for admixture in population genetics, shape recognition in computer vision, and testing for quasi-periodic signals in respiratory disease monitoring [1, 2, 3]. These applications and the broader theory and practice of topological data analysis begin with a point cloud of unlabeled data in some space [4]. To the best of our knowledge, there is no existing work on topologically analyzing data when the data points come with labels.

In this paper, we investigate data sets with class labels, in which the most important shape is the shape of the boundary between the classes. Known as the decision boundary, this boundary between classes is precisely what is learned by supervised classification algorithms [5]. The shape of the decision boundary, rather than the shape of the individual class-conditional point clouds, is what determines the complexity of a data set [6].

The primary topic in topologically studying data sets is *persistent homology*. The main steps in a persistent homology analysis are treating each data point as a node in a graph, connecting nearby nodes with edges where nearby is according to a scale parameter, forming complexes from the simplices formed by the nodes and edges, and examining the topology of the complexes as a function of the scale parameter. The topological features such as connected components, and holes and cavities of various dimensions that persist across scales are the ones that capture the underlying shape of the data set.

When we attempt to investigate the topology of the decision boundary of a labeled data set, we need to change the procedure for connecting nearby nodes and for forming complexes. All other parts of persistent homology analysis can remain unchanged. Specifically, since the decision boundary lies between the classes (subject to noise), we initially only connect nearby nodes of different labels with edges. However, such cross-label edges can only produce single points and line segments as simplices, not triangles, tetrahedrons, and so on. Therefore, we also introduce edges along walks of length 2 in the graph that start and end at nodes of different labels. With these additional edges, we can achieve higher order simplices and capture the topology of the decision boundary through persistent homology.

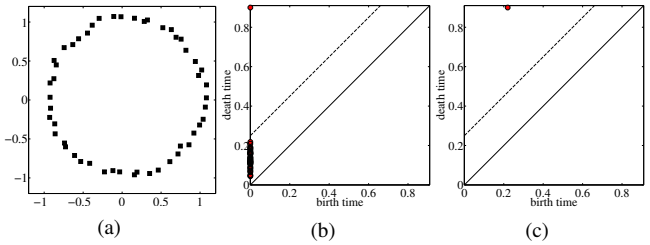
As mentioned earlier, decision boundaries arise most often in the context of classification problems. Many classification methods rely on kernel representations for obtaining nonlinear boundaries in the space of the input data set point cloud, but choosing the right kernel is often still a trial-and-error process. As one use case for topological analysis of decision boundaries, we examine the kernel selection problem. Specifically, we examine radial basis function (RBF) kernels and polynomial kernels in the context of support vector machine classifiers. We take advantage of the fact that the scale in persistent homology has an intimate relationship to the scale of the RBF. Also, since the decision boundary is an *algebraic variety* when polynomial kernels are used, we take advantage of relationships between the polynomial orders of algebraic varieties and quantifications of topological features [7].

The remainder of the paper is organized as follows. In Section 2, we provide background information on persistent homology of (unlabeled) data sets. In Section 3 we propose the modifications needed to the usual topological data analysis procedure in order to examine the shape of the decision boundary. Section 4 presents the application to kernel selection. We present empirical results on synthetic and real-world data sets in Section 5. Section 6 is the conclusion.

2. BACKGROUND ON PERSISTENT HOMOLOGY

Consider a set of T data points in $\mathbb{R}^n$: $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$. A set of points by itself has no shape per se, but if the points are viewed as samples from some shape, then the set of points reveals the underlying shape. We would like to estimate and approximate the topology of that shape by constructing a simplicial complex from the points and examining the topology of the simplicial complex. A zero-dimensional simplex is a point, a one-dimensional simplex is a line segment, a two-dimensional simplex is a triangle, a three-dimensional simplex is a tetrahedron, and so on; a simplicial complex is a set of simplices glued together in a particular way. Specifically, a simplicial complex $\mathcal{S} = (\mathcal{X}, \Sigma)$, where Σ is a family of non-empty subsets of $\mathcal{X}$ such that each subset $\sigma \in \Sigma$ is a simplex. Furthermore, the following condition must also hold: $\sigma \in \Sigma$ and $\tau \subseteq \sigma$ implies that $\tau \in \Sigma$. In forming these non-empty subsets of points that form a simplex, we only consider subsets of points that are close to each other. There are various notions of closeness that we come back to later in this section.

Topology, being the study of shape, is primarily concerned with the number of connected components and the number and dimen-



(a) (b) (c)

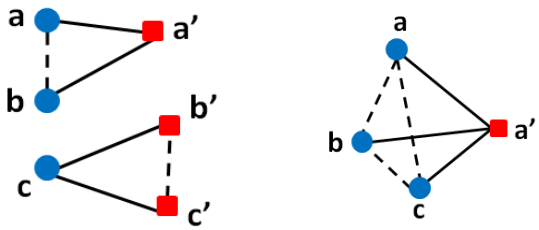
Fig. 1. (a) Noisy data samples from an underlying circle in 2-dimensional space, (b) persistence diagram for H_0, (c) persistence diagram for H_1.

sion of holes that an object has. The Betti numbers characterize the connectivity as follows. The zeroth Betti number β_0 is the number of connected components, the first Betti number β_1 is the number of one-dimensional holes or circles, the second Betti number β_2 is the number of two-dimensional voids or cavities, and so on. For example, a torus or inner tube has $\beta_0 = 1$ because it is just one component, $\beta_1 = 2$ because of the main hole through the middle and the hole formed when looking at a cross-section, and $\beta_2 = 1$ because of the cavity of the inner tube. Betti numbers for simplicial complexes are defined in the same way. Formally, $\beta_k(\mathcal{S})$ is the dimension of the kth homology group of the complex $H_k(\mathcal{S})$ [4].

Various approaches exist for constructing simplicial complexes from $\mathcal{X}$. All of these depend on a scale parameter ϵ (also referred to as *time*) which specifies the extent of closeness of points. In the Čech complex Čech$(\mathcal{X}, \epsilon)$, a simplex is created between a set of points $\mathcal{G}$ if and only if there is a non-empty intersection of the closed Euclidean balls $B(\mathbf{x}_i, \epsilon/2), \forall i \in \mathcal{G}$. In the Vietoris-Rips (VR) complex, VR$(\mathcal{X}, \epsilon)$, a simplex is created if and only if the Euclidean distance between every pair of points is less than ϵ. Efficient construction of the VR complex can proceed by creating an $\epsilon-$neighborhood graph, also referred to as the *one-skeleton* of $\mathcal{S}$. Then inductively, triplets of edges that form a triangle are taken as two-dimensional simplices, sets of four two-dimensional simplices that form a tetrahedron are taken as three-dimensional simplices, and so on.

Homological inference depends on the scale parameter (time) at which the complexes are constructed. The topological features of the simplicial complex $\mathcal{S}$ constructed from the data points $\mathcal{X}$ that are stable across scales, i.e. that are *persistent*, are the ones that provide information about the underlying shape. Topological features that do not persist are noise. *Persistence diagrams* are representations of the birth and death times of each homology cycle corresponding to each homology group H_k, $k = 0, 1, \ldots$, i.e. for increasing values of the scale parameter, the ϵ value at which a topological feature begins to exist and ceases to exist.

As an example, let us consider the point cloud $\mathcal{X}$ shown in Fig. 1(a), with noisy samples drawn from a circle, which has Betti numbers $\beta_0 = 1$, $\beta_1 = 1$, and $\beta_k = 0$ for $k > 1$. At the value $\epsilon = 0$, the simplicial complex that is formed from $\mathcal{X}$ is a collection of all the individual points not connected to any other point, resulting in the birth of T topological features in the H_0 persistence diagram shown in Fig. 1(b). As the scale increases, all of these little features die and only one persists until the largest scale under consideration; thus the persistent $\beta_0 = 1$. Looking at the H_1 persistence diagram in Fig. 1(c), we see that the only feature that is born persists until the largest scale and thus the persistent $\beta_1 = 1$. It is born at approximately a scale parameter of 0.2, which is when all of the points have been connected into a ring in the simplicial complex.



(a) (b)

Fig. 2. (a) A simplicial complex with two 2-simplices from a bipartite graph between circle and square classes generated using length-2 walks (dotted lines), (b) a complex created with one 3-simplex using the same approach.

3. SKELETONS AND SIMPLICIAL COMPLEXES FOR DECISION BOUNDARIES

As we discussed in the introduction, we are dealing with point clouds with labels. For simplicity, let us consider the binary case with the labels $y \in \{-1, +1\}$ so that we have the pairs $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_T, y_T)\}$. In order to understand the topology of the decision boundary that separates the classes $+1$ and -1, we will first consider neighborhood samples across the classes.[1]

We first connect nodes $\mathbf{x}_i$ and $\mathbf{x}_j$, $i \neq j$, if $y_i \neq y_j$, and $\mathbf{x}_i \in \mathcal{N}(\mathbf{x}_j)$ or $\mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i)$ [8]. One neighborhood we consider is the ϵ-neighborhood $\mathcal{N}_\epsilon$, in which case the neighborhood membership is symmetric. The result of this procedure is a bipartite graph between the classes with unweighted adjacency matrix $\mathbf{A}$. The diagonal of $\mathbf{A}$ is all zeroes.

In bipartite graphs, the only possible simplices are points and line segments. Higher-dimensional simplices such as triangles and tetrahedrons are not possible since there are no edges between interclass samples. To allow us to better capture the topology of the decision boundary, we would like to also include such higher-dimensional simplices. We also add edges arising from graph walks of length two, which introduces intraclass edges and therefore higher-dimensional simplices. To construct the adjacency matrix with the length two walks, which we denote $\tilde{\mathbf{A}}$, we first take $(\mathbf{A} + \mathbf{I})^2$ and then change all nonzero values to one to preserve the unweighted nature of the graph. To illustrate the procedure, we show two simple examples in Fig. 2. In the first example, we start with three points in a two-dimensional space where all points are within ϵ of each other. Two share a class label and are thus not initially connected by an edge. The initial graph $\mathbf{A}$ has two line segment simplices. After including the graph walk, an intraclass edge is introduced. Now $\tilde{\mathbf{A}}$ has a triangle simplex. The second example is similar, but has four points in three-dimensional space, with three of the four points sharing a class label. Here we form a tetrahedron after introducing the length two graph walk edges.

We define the graph encoded by $\tilde{\mathbf{A}}$ to be the one-skeleton of the simplicial complex. Hence, a simplex of dimension $(r+1)$ is inductively included in the simplicial complex if all of its r-dimensional faces are included. Using existing methods for unlabeled data sets, we can then calculate the persistence diagrams and Betti numbers from the resulting simplicial complexes at different scales.

[1]Multiclass extensions can consider the decision boundaries in one-against-one, one-against-all, and Venn diagram constructions [5].

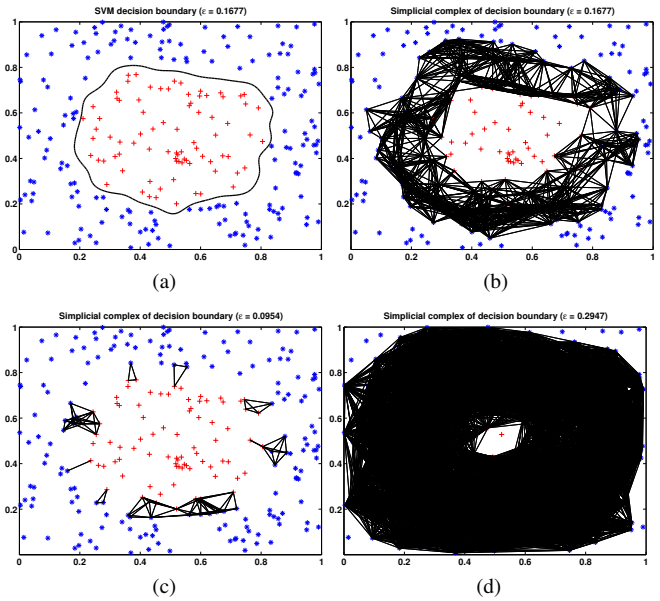


Fig. 3. (a) SVM decision boundary at optimal ϵ (RBF kernel), (b) one-skeleton at optimal ϵ inferred by our method, (c) one-skeleton at smaller ϵ, and (d) one-skeleton at larger ϵ.

4. APPLICATION TO KERNEL SELECTION

A common task when dealing with labeled data sets is learning a classifier that generalizes. Classifiers are defined by their decision boundaries, which are often represented as the zero level sets of kernel representations. Although there are existing methods for choosing a kernel, the process is still very much an art rather than a science. In this section, focusing on the Gaussian RBF kernel and the polynomial kernel, we comment on how topological data analysis of decision boundaries, as we have described in the paper thus far, can be used in the kernel selection process.[2]

First let us give the Gaussian RBF and polynomial kernel representations. The general form of a kernel representation is:

$$f(\mathbf{x}) = \sum_{i=1}^{T} \alpha_i k(\mathbf{x}_i, \mathbf{x}), \tag{1}$$

where α_i are coefficients. The set $\{\mathbf{x}|f(\mathbf{x}) = 0\}$ is the decision boundary and $\hat{y}(\mathbf{x}) = \text{sign}\left(f(\mathbf{x})\right)$ is the classifier. With the Gaussian RBF kernel, $k(\mathbf{x}_i, \mathbf{x})$ is $\exp\left(\|\mathbf{x}_i - \mathbf{x}\|^2 / \left(2\epsilon^2\right)\right)$ and with the polynomial kernel, $k(\mathbf{x}_i, \mathbf{x})$ is $\left(\mathbf{x}_i^T \mathbf{x} + 1\right)^d$.

The ϵ in the RBF kernel is a scale parameter akin to the scale parameter in persistent homology. We conjecture that values of ϵ at which topological features of the decision boundaries persist are also optimal values for the ϵ of the RBF kernel. We examine this further empirically in Section 5.

The decision boundary obtained from a polynomial kernel representation, i.e. its zero level set, is an algebraic variety. An algebraic variety is defined in general as the set of solutions of a system of polynomial equations. There exist relationships between Betti numbers, and the polynomial degree d and dimension of the space n of algebraic varieties in the literature [7]. One example of such a result is that if the polynomial degree is less than or equal to d, then [9, Thm. 2]:

$$\sum_{k=0}^{\infty} \beta_k \leq d(2d-1)^{n-1}. \tag{2}$$

The contraposition of this statement is that if the sum of the Betti numbers of the decision boundary is greater than $d(2d-1)^{n-1}$, we will require a polynomial kernel with degree greater than d. Thus we can select an appropriate polynomial degree through topological data analysis of the decision boundary. For example, if the decision boundary in a two-dimensional space is a circle, so that the sum of the Betti numbers is 2, we will need a polynomial order greater than 1.2807, i.e. at least a quadratic ($d = 2$) kernel after taking the ceiling, to be able to learn the decision boundary.

[2]Being the early stages of topological data analysis research, using persistent homology for kernel selection is not currently computationally competitive but may become so with further development.

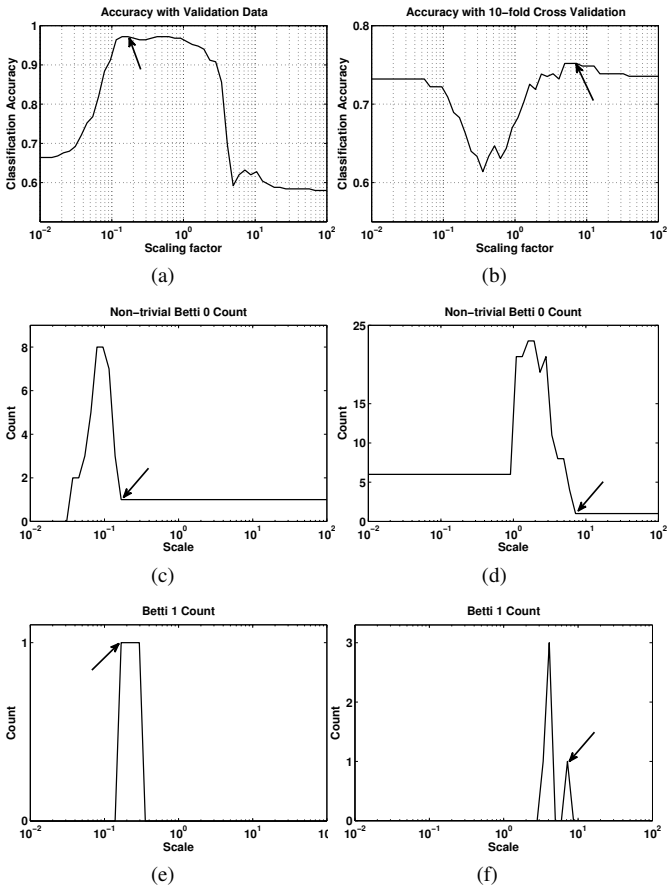


Fig. 4. (a)–(b) SVM accuracy, (c)–(d) non-trivial β_0, and (e)–(f) β_1 for *circle* (left) and *Haberman* (right). Arrows indicate optimal ϵ.

5. EMPIRICAL RESULTS

We perform topological analysis on decision boundaries of four data sets. In the first data set, *circle*, each point is sampled uniformly from $[0, 1]^2$ and labeled -1 if within a circle centered at $(0.5, 0.5)$ with radius 0.1, and labeled $+1$ otherwise. A noisy version of this construction, *noisy circle*, is created by randomly assigning a label to points in the annular region between radii 0.06 and 0.14. We create 250 samples each for training and validation in both data sets. The

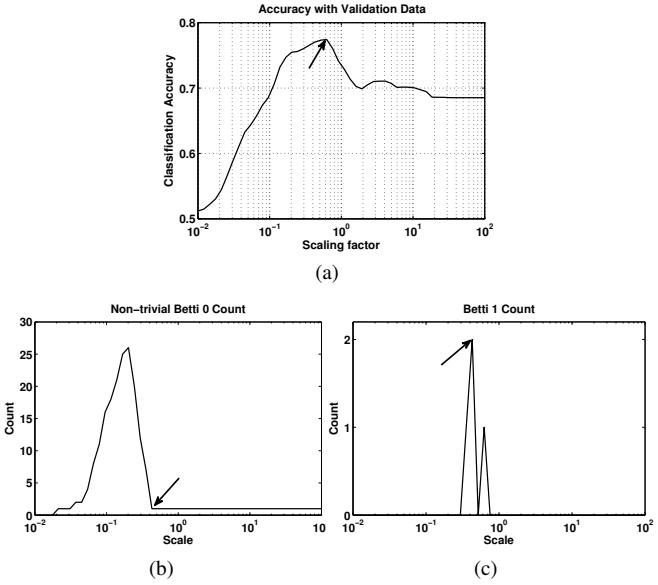


Fig. 5. (a) SVM accuracy, (b) non-trivial β_0, and (c) β_1 for *ESL mixture*. Arrows indicate optimal ϵ.

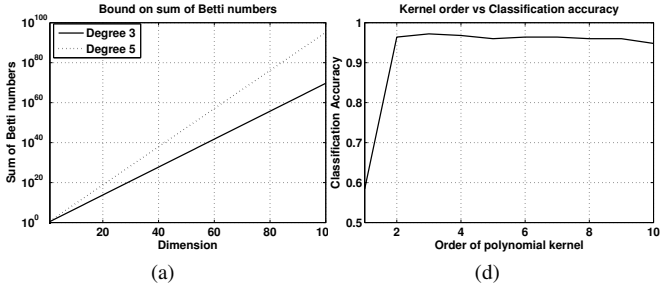


Fig. 6. (a) Upper bound on sum of Betti numbers for polynomials, (b) classification accuracy for *circle* with various polynomial degrees.

third data set, *Haberman* from the UCI machine learning repository [10], represents the survival of patients after breast cancer surgery and has 3 covariates and 306 samples. The fourth data set, *ESL mixture* from [11], has 100 samples per class; we also create a validation set for this data with $10,000$ samples using the generating distributions provided.

We first validate the ability of the proposed approach to infer the optimal ϵ for the RBF kernel. For *circle*, the optimal ϵ for SVM with RBF kernel is 0.1677 and the decision boundary is shown in Fig. 3(a); clearly it has $\beta_0 = 1$ and $\beta_1 = 1$. The simplicial complex at the same scale obtained using the proposed approach is shown in Fig. 3(b); it has the same Betti numbers. Furthermore, visual inspection shows that the simplicial construction approximates the decision boundary closely. Note that the '$+$'s and '$*$'s denote samples of the two classes and the lines that connect them indicate the one-skeleton from which the complex is formed. The constructions for two other scales ($\epsilon = 0.0954, 0.2957$) are shown in Figs. 3(c) and 3(d), demonstrating the progression of the complex from lower to higher scales. The total number of simplices are 313, 5690, and 137926 for the three ϵ values in increasing order.

The classification accuracies and Betti numbers for various ϵ values are shown for *circle* and *Haberman* in Fig. 4. The plotted count of β_0 excludes the contribution of zero-simplices of the complex (stand-alone samples), and hence truly represents the number of connected components in the decision boundary complex. The ϵ at which the non-trivial β_0 becomes constant after the initial rise and fall is the optimal scale chosen by our method. The reasoning behind this is that the non-trivial β_0 initially increases as the decision boundary complex is formed (Fig. 3(c)), but simplices eventually merge at some scale (Fig. 3(b)) to create stable features for the decision boundary complex. These features persist, e.g., in *circle* the Betti numbers $\{1, 1\}$ persist from $\epsilon = 0.1677$ to $\epsilon = 0.2947$, reflecting the true topology of the decision boundary.

In both data sets, the ϵ computed using the proposed procedure coincides with the optimal ϵ value for the kernel SVM. For *circle*, the optimal ϵ value is 0.1677 using both SVM validation data and our topological method. For *Haberman*, the optimal ϵ value is 7.197 using both tenfold cross-validation and our method. The non-trivial $\beta_0 = 1$ and $\beta_1 = 1$ at this ϵ. Note that β_1 does not persist for more than one scale value and can be construed as topological noise. Similarly for *noisy circle*, the optimal ϵ value is 0.1389 via both SVM validation and our approach. As in *circle*, the non-trivial $\beta_0 = 1$ and $\beta_1 = 1$ near the optimal scale. Finally, on *ESL mixture*, the optimal ϵ using SVM validation is 0.6251 and the proposed approach yields an optimal ϵ of 0.4292, yielding a loss in accuracy of just 0.0037. The Betti numbers obtained are $\{1, 1\}$ and again β_1 represents noise since it does not persist.

We consider determining the degree of polynomial kernels from the decision boundary's topology as the second application of our method. The upper bound on the sum of Betti numbers for various data dimensions is shown for several polynomial degrees in Fig. 6(a). For linear kernel ($d = 1$), the upper bound is 1 for all dimensions, which means that for a linear kernel to work, the decision boundary has to be a single connected component with no higher order topological features. The figure also shows that complex decision boundaries with many connected components and holes can be modeled using higher order polynomial kernels. In *circle*, $\sum_{i=1}^{\infty} \beta_i = 2$, and hence by (2), we need at least a second order kernel. The accuracies plotted in Fig. 6(b) justify this choice, since a degree 1 kernel does not achieve high accuracy but degree 2 and higher do.

6. CONCLUSION

We proposed a novel approach to construct simplicial complexes of decision boundaries in labeled data sets and estimate their topology through persistent homology. One application of this work is to prescribe kernels for classifying the labeled data set. This initial work opens the door to interesting future research directions in understanding and visualizing labeled data, and also in model selection for other classification approaches. Another possible research direction is investigating the effect of corrupted data, outliers, and random projection on estimates of decision boundary topology [12].

7. ACKNOWLEDGMENTS

The authors thank Henry Adams for answering JavaPlex questions and Visar Berisha for discussions.

8. REFERENCES

[1] M. Robinson, *Topological Signal Processing*. New York, NY: Springer, 2014.

[2] S. Emrani, T. Gentimis, and H. Krim, "Persistent homology of delay embeddings and its application to wheeze detection," *IEEE Signal Process. Lett.*, vol. 21, no. 4, pp. 459–463, Apr. 2014.

[3] D. Yorukoglu, F. Utro, D. Kuhn, S. Basu, and L. Parida, "Topological signatures for population admixture," in *Proc. Int. Conf. Res. Comput. Molec. Bio.*, Warsaw, Poland, Apr. 2015.

[4] G. Carlsson, "Topology and data," *Bull. Amer. Math. Soc.*, vol. 46, no. 2, pp. 255–308, Apr. 2009.

[5] K. R. Varshney and A. S. Willsky, "Classification using geometric level sets," *J. Mach. Learn. Res.*, vol. 11, pp. 491–516, Feb. 2010.

[6] T. K. Ho and M. Basu, "Complexity measures of supervised classification problems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 289–300, Mar. 2002.

[7] S. Basu, R. Pollack, and M.-F. Roy, "Betti number bounds, applications and algorithms," in *Combinatorial and Computational Geometry*, J. E. Goodman, J. Pach, and E. Welzl, Eds. New York, NY: Cambridge University Press, 2005, pp. 87–96.

[8] J. H. Friedman and L. C. Rafsky, "Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests," *Ann. Statist.*, vol. 7, no. 4, pp. 697–717, Jul. 1979.

[9] J. Milnor, "On the Betti numbers of real varieties," *Proc. Am. Math. Soc.*, vol. 15, no. 2, pp. 275–280, Apr. 1964.

[10] K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[11] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York, NY: Springer, 2009.

[12] K. N. Ramamurthy, K. R. Varshney, and J. J. Thiagarajan, "Computing persistent homology under random projection," in *Proc. IEEE Int. Workshop Stat. Signal Process.*, Gold Coast, Australia, Jun.–Jul. 2014, pp. 105–108.