Research Statement

Nakul Verma

Overview

The information explosion of the past few decades has created tremendous opportunities for Machine Learning-based data analysis. Such data typically possesses a large number of features. Take for instance, the task of predicting whether a particular treatment would be effective for a patient by analyzing their genome information. One hopes that by measuring several gene expressions, one can capture the necessary information about the data, leading to better predictions. However, the presence of a large collection of irrelevant features just add to the computational complexity of the learning algorithm, without helping much to solve the task at hand. Indeed, conventional statistical wisdom dictates that in a general setting the learning task becomes significantly more difficult with increase in the number of features. This makes it especially difficult to design and analyze learning algorithms for modern high-dimensional datasets.

I am broadly interested in different ways one can cope with this *curse of dimension*. The basic observation is that while modern datasets are represented in high dimensions, they often adhere to some low-dimensional *intrinsic* structure. This intrinsic structure can be manifested in data in several forms: some datasets such as text have a sparse structure, other datasets such as speech and image articulations follow a manifold structure, while other datasets such as genomics data typically conform to a taxonomy structure. If we can determine that the intrinsic structure is in fact low-dimensional (that is, has few degrees of freedom), then we can expect that the complexity of learning algorithms should scale only with data's intrinsic dimensionality.

My past and current work focuses on theoretical and practical issues in designing effective learning algorithms (both unsupervised and supervised) when the given data does adhere to some known underlying intrinsic structure (Sections 1 and 2). I will then discuss my work on understanding and formalizing the general phenomenon of *low intrinsic dimension* and its effects on the complexity of learning (Section 3).

1 Unsupervised and weakly-supervised methods for manifold structured data

Unsupervised dimensionality reduction of manifolds. Manifold dimensionality reduction has received considerable attention in the past decade. Given an n-dimensional manifold in $\mathbb{R}^D$, the goal is to come up with an embedding of the data in $\mathbb{R}^d$ ($d \ll D$) that preserves some interesting property of the manifold, say, preserve interpoint geodesic distances. This problem was originally studied in differential geometry by Nash [1954] and Kuiper [1955]. Their results show that one can embed the n-dimensional manifold in just $d = 2n + 1$ dimensions while preserving all geodesic distances. In the context of Machine Learning, we want to develop algorithms that achieve something similar when we have access to only a *finite size* sample from the underlying manifold.

While several algorithms are suggested to solve this problem, only a few provide any sort of guarantee on the quality of the embedding. My recent works [Verma, 2011, 2012] explore manifold embedding techniques that can provide good distance preserving guarantees. [Verma, 2011] shows that a random linear projection of a general n-dimensional manifold into $\tilde{O}(n/\epsilon^2)$ dimensions can

preserve lengths of all paths on a manifold (not just the geodesics) within a factor of $(1 \pm \epsilon)$. This paper provides an alternate technique to analyze random projections for manifolds, and improves upon the previous known results by Baraniuk and Wakin [2007] and Clarkson [2007].

My second work [Verma, 2012] tackles the problem of minimizing the dependence on ϵ that is typically associated with random projection type results. Following the ideas discussed by Nash [1954], we can show that certain non-linear mappings can *completely remove* the dependence on ϵ from the embedding dimension. We derive an explicit algorithm for computing the target embedding based on samples from the underlying manifold and show that our embedding approximately maintains the geodesic distances between any pair of points from the underlying manifold (not just the input samples). This work can be viewed as an algorithmic realization of Nash's Embedding Theorem, and provides the sharpest known result for algorithmically embedding arbitrary n-dimensional manifolds in just $\tilde{O}(n)$ dimensions while maintaining the underlying geodesic distances.

My results on manifold dimensionality reduction validate our intuition that the complexity (in this case, the dimension of the embedding space) of learning depends only on the dimension of the intrinsic structure of the data (the dimension of the underlying manifold).

Sample complexity of Multiple Instance Learning (MIL) for manifold bags. MIL is a popular weakly-supervised learning paradigm in which training data is provided in the form of labeled *sets of instances* (or bags) rather than individual instances. Typically, a bag is labeled positive if some instance in the bag is positive. My work [Babenko et al., 2011] reconciles the disconnect between how MIL is used in practice and its theoretical analysis. We argue that in many applications of MIL the data is better modeled as low dimensional manifold bags in high dimensional feature space, and show that the geometric properties of such manifold bags is intimately related to its PAC-learnability. We also present a practical algorithm for MIL using manifold bags and show that it works well for real-world detection tasks in image and audio domains.

These results corroborate our intuition that even though data may be represented in high dimensions, if the data distribution conforms to a low-dimensional manifold structure, then the learning complexity does in fact scale with the complexity of the underlying manifold and not the ambient space.

Future directions. An obvious next direction is to explore whether something similar can be shown for other types of popular intrinsic structures, say, a cluster structure. Researchers have extensively studied unsupervised learning of individual components—or clusters—of data generated from a mixture distribution. Most works focus on analyzing the case when the individual clusters follow a Gaussian or a log-concave distribution.

In view of modern datasets, it may be more appropriate to assume that the dataset is a mixture distribution where the individual components are distributions that factor over a generic Bayes Net. Learning such a structure would help us capture higher order relationships within each cluster, resulting in a better model that can predict well from our underlying data distribution. I plan to investigate a scheme whereby one can learn such a mixture by (i) doing a non-linear dimension reduction that makes the individual components more pronounced, followed by (ii) using an EM-type algorithm to recover the individual components.

2 Exploiting intrinsic structure for supervised learning

My work in supervised learning shows strong empirical evidence for low-dimensional structure for individual application domains.

Metric learning for taxonomy structured data. Modern multi-class datasets are often part of an underlying semantic taxonomy. My recent work [Verma et al., 2012] shows that one can leverage this taxonomy structure to learn better similarity metrics for the input data. We show that a nearest

neighbor classifier using the learned metrics gets improved performance over the best discriminative methods. Moreover, our learned metrics can also help in some taxonomy specific applications. We show that the metrics can help determine the correct placement of a novel category that was not part of the original taxonomy, and can provide effective classification amongst categories local to specific subtrees of the taxonomy. Experimental analysis shows that the class discrimination information in the learned metrics is encoded in the top few principal components, thus exhibiting strong evidence of the few degrees of freedom phenomenon.

Low-dimensional modeling of spatio-temporal data. Many problems in the sensor networks involve collecting data across a large physical region over time. Since nodes in a sensor network experience periodic downtimes due to limited energy, recovering missing sensor data is a critical problem. Fortunately, such data is often highly correlated across space and time forming a *spatio-temporal structure*. My work [Verma et al., 2011] explores how well one can exploit this structure by using latent variables. We show that low-dimensional latent variables can effectively reconstruct missing sensor data values under various practical downtime scenarios. The effectiveness of low-dimensional latent variables in reconstructing missing data gives a strong indication that such spatio-temporal structure is also low dimensional.

Future directions. It is interesting to explore what one can do if we only have *partial information* about the intrinsic structure. Taking the taxonomy structure for instance, how is the quality of classification affected when one only has a corrupted version of the taxonomy? Going a step further, it is worth investigating whether we can *fix* the corrupted taxonomy. I plan to investigate how one can use a non-parametric Bayesian method like a nested Chinese Restaurant Process to induce a *prior* over the space of taxonomies. Such a prior would help mitigate the effects of the corrupted taxonomy, resulting in improved performance.

3 Formalizing low-dimensional intrinsic structure and its implications on learning

Since low-dimensional intrinsic structure is prevalent in modern datasets, one of the key problems is how can one quantify these intrinsic degrees of freedom of the underlying data distribution *without knowing the exact nature of the intrinsic structure*. Such a characterization should be (i) conducive to algorithmic analysis (that is, be able to provide guarantees for the learning algorithm), (ii) robust to noise, and (iii) empirically verifiable from a finite size sample. While certain notions of characterizing intrinsic complexity such as *doubling dimension* and *covering dimension* do seem to satisfy requirement (i) and, with certain modifications to the definitions, (ii), the biggest hurdle seems to be how can one empirically verify that real-world datasets do in fact have low, say, doubling dimension.

My work [Verma et al., 2009] explores to what extent real-world datasets adhere to an alternate notion of characterizing intrinsic complexity called the *local covariance dimension* (formalized in [Freund et al., 2007, Dasgupta and Freund, 2008]). We show that certain datasets do in fact have low intrinsic complexity (in terms of their covariance dimension), and the learning complexity of tree based space partitioning algorithms is closely tied with this quantity. Extensive experimental analysis shows that the quality of regression estimates, nearest neighbor queries and quantization errors yielded by these space partition algorithms scales with the local covariance dimension of the underlying data distribution.

Current and future directions. Developing on this theme, I am currently working on a project that provides formal statistical guarantees on estimating other notions of intrinsic dimension when one has access to only a finite size sample. Without assuming a specific intrinsic structure, we are exploring to what extent this phenomenon of low degrees of freedom is prevalent in datasets from a

wide range of application domains. Having statistically sound estimates of the intrinsic dimension would help us better understand the complexity of learning on different datasets, and would enable us to determine *a priori* how well will certain learning algorithms perform on a given dataset.

References

B. Babenko, N. Verma, P. Dollar, and S. Belongie. Multiple instance learning with manifold bags. *International Conference on Machine Learning (ICML)*, 2011.

R. Baraniuk and M. Wakin. Random projections of smooth manifolds. *Foundations of Computational Mathematics (FoCM)*, 2007.

K. Clarkson. Tighter bounds for random projections of manifolds. *Comp. Geometry*, 2007.

S. Dasgupta and Y. Freund. Random projection trees and low dimensional manifolds. *ACM Symposium on Theory of Computing (STOC)*, 2008.

Y. Freund, S. Dasgupta, M. Kabra, and N. Verma. Learning the structure of manifolds using random projections. *Neural Information Processing Systems (NIPS)*, 2007.

N. Kuiper. On C^1-isometric embeddings, I, II. *Indag. Math.*, 17:545–556, 683–689, 1955.

J. Nash. C^1 isometric imbeddings. *Annals of Mathematics*, 60(3):383–396, 1954.

N. Verma. A note on random projections for preserving paths on a manifold. *UC San Diego, Tech. Report CS2011-0971*, 2011.

N. Verma. Distance preserving embeddings for general n-dimensional manifolds. *Under Review (JMLR)*, 2012.

N. Verma, S. Kpotufe, and S. Dasgupta. Which spatial partition trees are adaptive to intrinsic dimension? *Uncertainty in Artificial Intelligence (UAI)*, 2009.

N. Verma, P. Zappi, and T. Rosing. Latent variables based data estimation for sensing applications. *International Conference on Intelligent Sensors, Sensor Networks, and Information Processing (ISSNIP)*, 2011.

N. Verma, D. Mahajan, S. Sellamanickam, and V. Nair. Learning hierarchical similarity metrics. *Computer Vision and Pattern Recognition (CVPR)*, 2012.