High-Dimensional Data Processing using Sample Compression and Dimensionality Reduction

Many modern applications of computing, involve the storage and processing of vast amounts of high-dimensional data. Designing algorithms and data structures to handle high-volume, high-dimensional data efficiently, is thus a key challenge faced by algorithm designers and programmers today.

For instance, consider an image search application which takes an image as input and returns the "closest match" from a database of images. Supposing each image is 48 megapixel and truecolor, this means that a single image consists of 48 million pixels, where each pixel has one of 16 million possible color values. Thus, each image is a point in a 48 million dimensional space, and image search simply means searching for the nearest neighbour of the input point (i.e. image). Even though in practice, the image is typically stored after preprocessing to extract a smaller number of "features", their number can still run into hundreds or thousands, which means the algorithm still has to work in several hundred dimensions.

An important technique to address these issues is the idea of *preprocessing*, that is, given a high-volume, high-dimensional dataset, we want to reduce its size and / or its dimensionality before presenting it to a data processing algorithm. Thus, we can broadly split such preprocessing techniques into two classes – those that reduce the *size* of the dataset or *sample compression techniques*, and those that reduce the *dimensionality*, or *dimensionality reduction* techniques. Naturally, in either case we would like such a reduction to preserve the information content of the original data as much as possible.

Such preprocessing techniques have been in use since the 80's and newer ones are being discovered more and more as computing technology advances. However, most of these techniques were designed to preserve information such as interpoint euclidean distances or certain notions of nearness / distance between points. Meanwhile in the last two decades, the wide-ranging success of neural networks, machine learning and artificial intelligence algorithms has given rise to new tools being used for data analysis, such as statistical distance functions and their generalizations known as *dissimilarity measures*. Another interesting development is that *higher-order* geometric properties of the data points have gained in importance. Many conventional sample compression and dimensionality reduction methods have unknown compatibility with these measures or do not work with them.

We aim to analyze existing preprocessing techniques with respect to such measures and geometric properties, as well as design new techniques which will work with them if they are not compatible with the existing techniques, and give mathematically proven guarantees over their functioning. For example, in the case of methods such as random projections for dimensionality reduction, their behaviour with respect to higher order geometric features, is not well-understood. Moreover, they work only with euclidean distances. Our goal shall be to analyze their behaviour with respect to higher-order geometric features, and devise new dimensionality reduction techniques that can be used with dissimilarity measures. Similarly in sample compression, the theory of random sampling based on the *VC* or Vapnik-Chervonenkis dimension, is fairly well-understood. Our goal shall be to adapt the VC theory to work with new dissimilarity measures, and to build on other techniques from topology and discrete geometry, to address cases resistant to the VC dimension approach. Lastly, *discrepancy minimization* is an alternate method to obtain sparse samples via iterated balanced partitions. We shall look to design improved algorithms for discrepancy minimization, in order to obtain sparse samples.

The primary results of the project shall be (i) new techniques for dimensionality reduction, (ii) mathematical guarantees on the effectiveness of known techniques with respect to new types of informations stored in the data. Moreover, we shall also look to generate (iii) improved sample compression bounds, with for systems satisfying the *bounded VC dimension* conditions, as well as (iv) new techniques of sample compression for systems with unbounded VC dimension, using mathematical tools from discrete geometry, topology and combinatorics. An algorithmic goal shall be (v) improved randomized or deterministic algorithms for discrepancy minimization. An expected result from the project is the building of bridges between high-dimensional algorithms on the one hand, and between different branches of theory, such as randomized algorithms, high-dimensional probability, compressed sensing and geometric analysis, on the other.