The Curse of Length

Aaron Hui, Byron Gao
Cornell University, Texas State University

Abstract

The nearest neighbor problem is one that comes up in a variety of applications, from searching to clustering. However, in 1999, Beyer demonstrated that under certain conditions with high dimensionality, the nearest neighbor problem was not a meaningful question to ask. It gave form to the “curse of dimensionality” that many researchers had described before. This influential work spawned a series of investigations of this concentration phenomenon, which, for the most part, was limited to vector spaces. In this paper, we extend this investigation to sequence spaces under edit distance and longest common subsequence measures, which do not have an inherent notion of dimension. We prove results under which sequences will concentrate, and conduct experiments on synthetic data that demonstrate situations where sequences do and do not concentrate as the length of the sequences go to infinity. Rather than the curse of dimensionality, it is a curse of length.

Introduction

We investigate concentration effects of distributions under the edit distance and longest common subsequence (LCS) measures. By concentration, we mean that as the sequence length tends to infinity, the ratio of the furthest point to the nearest point in the dataset relative to the query point tends to 1.

Edit Distance: The minimum number of insertions, deletions, and substitutions to transform one sequence into another. For example, kitten and sitten have an edit distance of 1.

Longest common subsequence: The length of the longest common subsequence between 2 sequences. For example, abcdef and ace have an LCS measure of 3 since the longest common subsequence is ace. This is different from the longest common subsequence, where characters must be next to each other.

Beyer’s Concentration Condition

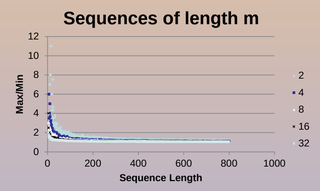
$$\lim_{m\to\infty} \frac{\mathrm{Var}[d(P_a, Q_m)]}{E[d(P_a, Q_m)]^2} = 0$$

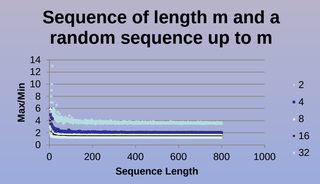
Theoretical Results

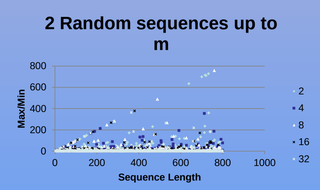
- Strings up to length m that are uniformly distributed concentrate under both edit distance and LCS
- A constant length string and a string up to length m that is uniformly distributed concentrate under edit distance

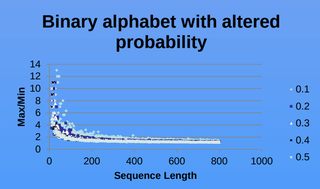
Experimental Results

Edit Distance

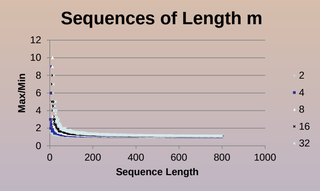


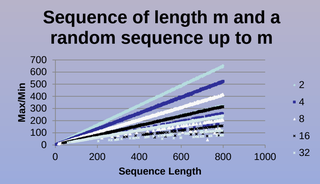


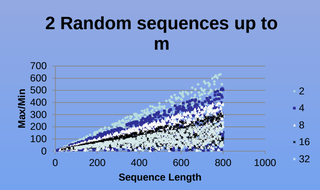


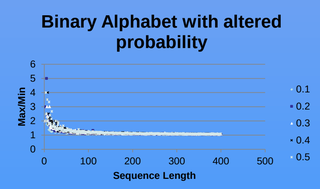


LCS









Conclusion

We proved that for distributions where each sequence has equal probability of occurring, they concentrate as the length of the sequence increases. From the experiments, it seems that as the distribution becomes farther from uniform, the distributions stop concentrating. Intuitively, this occurs due to the topology induced by edit distance and LCS on the space of sequences. Around a specific query point, as the maximum sequence length increases, the average distance of the query tends to increase fairly rapidly, due to the exponential growth of the number of sequences with sequence length. This leads to the observed concentration effect.

In this context, the length in the space of sequences behaves similarly to the dimension in vector spaces. However, it is also highly dependent on the topology induced by the distance function used. This seems to imply a more general situation, where increased data complexity tends to induce distance concentration. Intuitively, length and dimension seem to be important attributes of complexity. But there are many other data types, such as attribute data or structured data, that do not necessarily have a notion of dimension or length. The definition of data complexity is still somewhat vague, and the conditions under which “complexity” induces concentration requires more research.

The nearest neighbor problem for sequential data mining, and even for other metric spaces that cannot be easily described with vectors, deserves a closer look. The work done here is only a small step in that direction.

In the future, we plan to submit this to SIGMOD 2013.

Acknowledgements

This research is funded by the NSF REU Program