Approximate vs. Representative Nearest Neighbors

Thomas M. Breuel
U. Kaiserslautern and DFKI
Germany
tmb@informatik.uni-kl.de

Over the last several years, there has been renewed interest in efficient nearest neighbor search algorithms. Such algorithms have uses in areas like information retrieval, pattern recognition, data mining, compression, and databases. Known algorithms for the exact nearest neighbor problem have such complexities that, in practice, high dimensional nearest neighbor problems are usually solved with brute force search, that is, computing the distance of the query point with each data point in the database. In order to achieve better performance than brute force search, a large number of authors have considered the *approximate nearest neighbor* problem [4, 1, 3]. That is, instead of returning the exact nearest neighbor p to some query point q, they return, for a pre-specified approximation constant, an arbitrary neighbor p', such that $d(q, p') \leq (1 + \epsilon)d(q, p)$.

Approximation algorithms like these have been successful in finding practically useful solutions to a number of otherwise computationally hard problems. In many settings, a solution that is ϵ-approximate also incurs a extra cost proportional to ϵ. However, in the case of approximate nearest neighbor algorithms, things are not so clear-cut. Authors that have suggested the use of approximate nearest neighbor algorithms have often simply assumed that an approximate solution is "good enough" in many applications but have generally not provided a formal justification.

Nearest neighbor algorithms have some uses in which cost of a solution is proportional to the distance, and approximations therefore affect the cost of a solution predictably. However, when used in a decision theoretic context (e.g., pattern classification, regression, data mining, etc.), we show that this relationship does not hold anymore. In particular, while for exact nearest neighbor algorithms, the asymptotic classification error is known to be a small multiple of the Bayes-optimal rate, in the case of approximate nearest neighbor algorithms as commonly formulated in the literature, we can show that the classification error can become arbitrarily bad even for fixed ϵ and "well behaved" sample distributions. Fixing this problem requires modifications to the definition of an approximate nearest neighbor algorithm to include a notion of fair sampling of the candidate neighbors. We briefly discuss to what degree some of the existing approximate nearest neighbor algorithms are subject to such problems.

Second, we examine the sample distributions on which (approximate) nearest neigh-

bor methods are usually successful for decision theoretic problems. We observe that sample distributions on which nearest neighbor methods are useful (both empirically and theoretically) tend to have additional structure that can be exploited, resulting in better decision theoretic performance (in addition to the potential for improved running times).

Based on these observations, we introduce a class of algorithms that we call k-*representative neighbor algorithms*. These algorithms are based on a notion *statistically representative samples*. These algorithms behave like k-nearest neighbor algorithms, that is, they accept a set of training samples and a query point and return a set of k related points (though not necessarily nearest neighbors). Unlike approximate nearest neighbor methods, we can show that representative neighbor methods preserve the decision theoretic asymptotic bounds associated with exact nearest neighbor algorithms and that they have desirable non-asymptotic properties. Implementations of representative neighbor algorithms in terms of standard unsupervised learning methods (e.g., [2]) are given and algorithmic and computational complexity issues related to k-representative neighbor algorithms are discussed.

References

[1] Sunil Arya, David M. Mount, Nathan S. Netanyahu, Ruth Silverman, and Angela Y. Wu. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM*, 45(6):891–923, 1998.

[2] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd ed)*. Wiley Interscience, 2001.

[3] P. Indyk. *Handbook of Discrete and Computational Geometry, Second Edition*, chapter Nearest Neighbors in High-dimensional Spaces. CRC Press LLC, Boca Raton, FL, 2004.

[4] Jon M. Kleinberg. Two algorithms for nearest-neighbor search in high dimensions. In *Proc. 29th ACM Symposium on Theory of Computing, 1997*, pages 599–608, 1997.