Summary: In this paper, we consider the problem of classification of high dimensional queries to high dimensional classes from discrete alphabets where the probabilistic model that relates data to the classes is known. This problem has applications in various fields including the database search problem in mass spectrometry. The problem is analogous to the nearest neighbor search problem, where the goal is to find the data point in a database that is the most similar to a query point. The state of the art method for solving an approximate version of the nearest neighbor search problem in high dimensions is locality sensitive hashing (LSH). LSH is based on designing hash functions that map near points to the same buckets with a probability higher than random (far) points. To solve our high dimensional classification problem, we introduce distribution sensitive hashes that map jointly generated pairs to the same bucket with probability higher than random pairs. We design distribution sensitive hashes using a forest of decision trees and we analytically derive the complexity of search. We further show that the proposed hashes perform faster than state of the art approximate nearest neighbor search methods for a range of probability distributions, in both theory and simulations. Finally, we apply our method to the spectral library search problem in mass spectrometry, and show that it is an order of magnitude faster than the state of the art methods.

MSC:
68P15 Database theory
62H30 Classification and discrimination; cluster analysis (statistical aspects)

Keywords:
classification; discrete probability distribution; hash design; high-dimensional data; locality sensitive hashing

Full Text: DOI arXiv

References:


This reference list is based on information provided by the publisher or from digital mathematics libraries. Its items are heuristically matched to zbMATH identifiers and may contain data conversion errors. It attempts to reflect the references listed in the original paper as accurately as possible without claiming the completeness or perfect precision of the matching.