Journal of Complexity 25 (2009) 385–397

Contents lists available at ScienceDirect

Journal of Complexity

journal homepage: www.elsevier.com/locate/jco

When is 'nearest neighbour' meaningful: A converse theorem and implications

Robert J. Durrant, Ata Kabán *

School of Computer Science, The University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

ARTICLE INFO

Article history:
Received 1 July 2008
Accepted 28 February 2009
Available online 24 March 2009

Keywords:
High dimensionality
Distance concentration
Latent variable models

ABSTRACT

Beyer et al. gave a sufficient condition for the high dimensional phenomenon known as the concentration of distances. Their work has pinpointed serious problems due to nearest neighbours not being meaningful in high dimensions. Here we establish the converse of their result, in order to answer the question as to when nearest neighbour *is* still meaningful in arbitrarily high dimensions. We then show for a class of realistic data distributions having non-i.i.d. dimensions, namely the family of linear latent variable models, that the Euclidean distance will not concentrate as long as the amount of 'relevant' dimensions grows no slower than the overall data dimensions. This condition is, of course, often not met in practice. After numerically validating our findings, we examine real data situations in two different areas (text-based document collections and gene expression arrays), which suggest that the presence or absence of distance concentration in high dimensional problems plays a role in making the data hard or easy to work with.

1. Introduction

In an influential paper, Beyer et al. [1] point out a serious threat for indexing and similarity-based retrieval in high dimensional databases, due to the following phenomenon, called the concentration of distances: As the dimensionality of the data space grows, the distance to the nearest point approaches the distance to the farthest one. Nearest neighbours become meaningless. The underlying geometry of this phenomenon was further studied in [2], strongly suggesting the detrimental effects often termed informally as the 'curse of dimensionality' are attributable to this phenomenon.

* Corresponding author.
E-mail addresses: R.J.Durrant@cs.bham.ac.uk (R.J. Durrant), A.Kaban@cs.bham.ac.uk (A. Kabán).

0885-064X/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jco.2009.02.011

Beyond exponentially slowing down data retrieval [2], the problem of distance concentration is becoming a major concern more generally for high dimensional multivariate data analysis, and risks to compromise our ability to extract meaningful information from volumes of data [3,4]. This is because in many domains of science and engineering, the dimensionality of real data sets grows very quickly, while all data processing and analysis techniques routinely rely on the use of some notion of distance [4]. In particular, high impact application areas, such as cancer research, produce simultaneous measurements of the order of several thousands. As pointed out in [3], currently existing multivariate data analysis techniques were not designed with an awareness of such counter-intuitive phenomena intrinsic to very high dimensions. It is therefore imperative for this problem to be studied and better understood in its own right, before one can embark on trying to devise more appropriate computational techniques for high dimensional problems.

Despite its title "When is nearest neighbour meaningful" [1], the paper in fact answers a different question, namely "When nearest neighbour is not meaningful". In formal terms, they give a sufficient condition for the concentration phenomenon. However, knowing the answer to the previous question would be very important and useful, since then one would have an objective to work towards in order to get round of the problem, in principle. This is what we address in this paper.

Although many previous authors mention, and admit on the basis of empirical evidence, that cases exist when the nearest neighbour is still meaningful in high dimensions [5,1,4], generally valid formal conditions are still lacking. All recent formal analyses have been conducted assuming data distributions with i.i.d. dimensions [6,4], which is unrealistic in most real settings. Yet, it has been observed that, if techniques for mitigating the concentration phenomenon are used carelessly, they may actually end up having a detrimental effect [4].

Here we make the following contributions: We establish the converse of Beyer et al.'s result, which gives us a generic answer to when nearest neighbour *is* meaningful in arbitrarily high dimensions. Then, we give a class of examples of realistic data distributions having non-i.i.d. dimensions, where we show the Euclidean distance will not concentrate when the dimensionality increases without bounds, as long as the amount of 'relevant' dimensions grows no slower than the overall data dimensions. Of course, this condition is not always met in practice; examples will follow later.

These results provide a formal explanation for previous informal and empirical observations, such as [5] "increasing the input space dimension without enhancing the quantity of available information reduces the model's power and may give rise to the curse of dimension". Our theoretical result also provides a generic criterion that may be used as an objective to work towards in order to counter the problem when necessary.

2. Distance concentration

Let F_m, $m = 1, 2, \ldots$ be an infinite sequence of data distributions and $\boldsymbol{x}_1^{(m)}, \ldots, \boldsymbol{x}_n^{(m)}$ a random sample of n independent data vectors distributed as F_m. An arbitrary random vector distributed as F_m will be referred to as $\boldsymbol{x}^{(m)}$. For each m, let $\|\cdot\| : \mathrm{dom}(F_m) \to \mathbb{R}^+$ be a function that takes a point from the domain of F_m and returns a positive real value. Further, $p > 0$ will denote an arbitrary positive constant, and it is assumed that $\mathrm{E}[\|\boldsymbol{x}^{(m)}\|^p]$ and $\mathrm{Var}[\|\boldsymbol{x}^{(m)}\|^p]$ are finite and $\mathrm{E}[\|\boldsymbol{x}^{(m)}\|^p] \neq 0$ throughout this section.

In the context of the problem at hand, the interpretation of the function $\|\cdot\|$ is that of a distance metric (or norm)—though the theory does not rely on this interpretation, i.e. there is no requirement for it to satisfy the properties of a metric. Similarly, the positive integer m may be interpreted as the dimensionality of the data space.

Theorem 1 (*Beyer et al. [1]*). *If* $\lim_{m\to\infty} \frac{\mathrm{Var}[\|\boldsymbol{x}^{(m)}\|^p]}{\mathrm{E}[\|\boldsymbol{x}^{(m)}\|^p]^2} = 0$, *then* $\forall \epsilon > 0$, $\lim_{m\to\infty} P[\max_{1\leq j\leq n} \|\boldsymbol{x}_j^{(m)}\| < (1+\epsilon) \min_{1\leq j\leq n} \|\boldsymbol{x}_j^{(m)}\|] = 1$; *where the operators* $\mathrm{E}[\cdot]$ *and* $\mathrm{Var}[\cdot]$ *refer to the theoretical expectation and variance of the distributions* F_m, *and the probability on the r.h.s. is over the random sample of size n drawn from* F_m.

The proof can be found in [1].