February 16, 2019

The Blessings of Dimensionality: Theoretical analysis of nearest-neighbor projected-distance methods for finding interactions in high dimension

Bryan A. Dawkins[1], Trang T. Le[2] and Brett A. McKinney[1,3*]

[1]Department of Mathematics, University of Tulsa, Tulsa, OK 74104
[2]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104
[3]Tandy School of Computer Science, University of Tulsa, Tulsa, OK 74104

February 16, 2019

Abstract

It is commonly known that high-throughput data has many inherent statistical challenges, such as multiple testing, sparsity and over fitting. Collectively these challenges are known as the Curse of Dimensionality. Here we highlight an important Blessing of Dimensionality: the ability to identify interactions with nearest neighborhoods. We review nearest-neighbor concepts for finding interactions, and we derive important distribution moments for distance metrics in high dimensional spaces. We use these theoretical results and simulated data to offer recommendations for computational approaches to find nearest neighbors in high dimension. We discuss ways to maximize the blessings and minimize the curses of dimensionality to reliably identify interactions.

1 Introduction

Relief-based methods identify interacting attributes as important by using nearest-neighbor information in higher dimensions (the "blessings of dimensionality"). Myopic methods, such as univariate tests, that do not account for information from higher dimensions, are susceptible to false negatives when there are interactions. For example in the plot of variable A versus C in a three-variable simulation (Fig. 1a), variable A appears to show no difference between cases and controls (the marginal group means are the same). However, A is actually simulated to have a strong differential correlation with B, conditioned on the outcome variable (Fig. 2b). Current Relief-based methods determine the importance of an attribute by computing the average difference of a target instance (X) and its nearest instance form the same class (Hit) projected onto the attribute A dimension ($d_{\mathrm{X,H}}(A)$) subtracted from the projected difference of target X and its nearest instance from the opposite class (Miss) ($d_{\mathrm{X,M}}(A)$). When the inequality $d_{\mathrm{X,M}}(A) > d_{\mathrm{X,H}}(A)$, it suggests that attribute A is useful for discriminating between cases and controls.

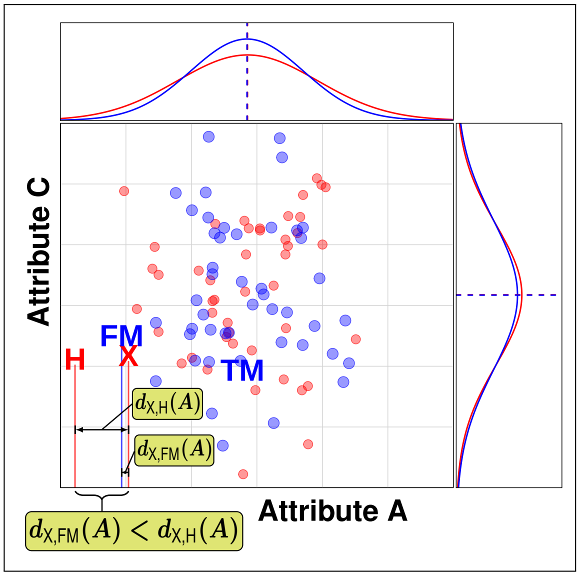

Fig 1. Imposters vs true neighbors in the presence of interactions with three variables. Scatter plot of simulated irrelevant Attribute C with a functional Attribute A **(a)**. None of the attributes has a main effect, but Attribute B and C interact through differential correlation **(b)**. Computing nearest neighbors with irrelevant attributes **(a)** or lower dimensions leads to imposter nearest neighbors and degrades the ability of Relief-based methods to identify interaction effects. Computing distances in only these two dimensions leads to an imposter false miss (FM) for the nearest neighbor from the opposite outcome class for target instance X. This imposter leads to attribute A predicting closer projected distances for misses than hits (H), which incorrectly indicates that A is a poor discriminator (yellow boxes in **(a)**). Computing nearest neighbors in higher dimensions **(c-d)** or with the correct interaction partner leads to imposter nearest neighbor (FM) being replaced by the true nearest miss neighbor (TM) for target instance X, which correctly leads to attribute A predicting closer projected distances for hits (H) than misses, which is an indication that attribute A is a good discriminator (yellow boxes **(b)**).

Relief-based methods use information from all attributes available to it (omnigenic) to estimate an attribute's importance. However, if relevant higher-dimensional information is not used, even Relief-based methods will miss the effect of A because "imposter" neighbors will be used in the attribute estimate (False Miss (FM) in Fig. 1, where $d_{\text{X,FM}}(A) < d_{\text{X,H}}(A)$). If one were to compute nearest neighbors in the A-C plane (ignoring the B dimension), the nearest miss would be an imposter (FM), which leads to a negative contribution to the importance score for A. One might call this C attribute a type-I confounding attribute because it increases the chances of interacting attributes to be false negatives. When nearest neighbors are calculated based on higher dimensions with relevant information (Fig. 2c), it is clear that TM is closer to X than FM. The imposter (FM) is replaced by the true nearest miss (TM) and attribute A correctly shows a greater projected difference between misses than hits (Fig. 2d $d_{\text{X,TM}}(A) > d_{\text{X,H}}(A)$), which is the signature of an important attribute. Univariate methods still cannot find the importance of A unless the interaction is explicitly modeled, but as long as functional variables A and B are in the space for nearest neighbor calculations (Fig. 2c-d), imposters can be excluded and Relief-based methods will find that A (and B) are important discriminators.

Using same interaction, increase background noise genes to see degrading of A and B Relief importance because of curse of dimensionality (sparseness).

1.1 Neighborhood Methods

Fixed-k, SURF, multiSURF, gene-wise adative-k.

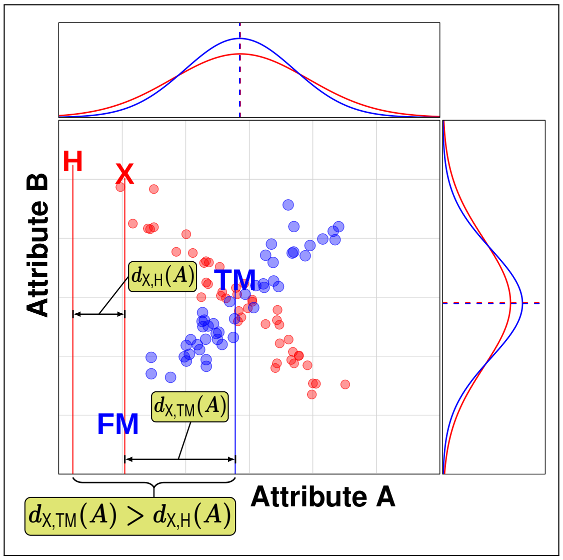

Fig 2. True neighbors

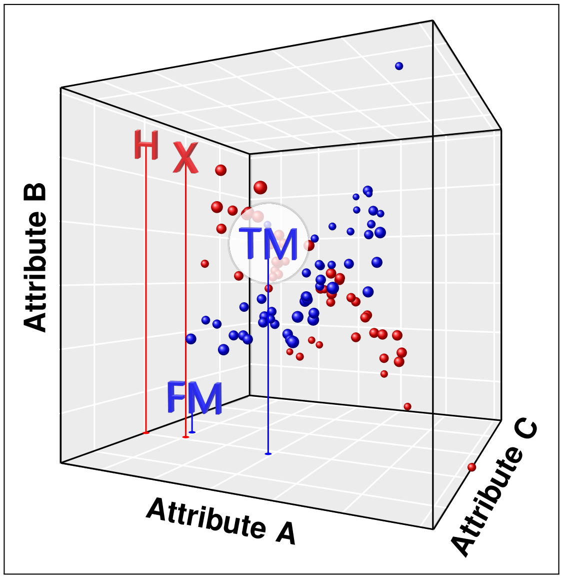

Fig 3. 3D AB view. Still working on this.

2 Derivation of expected k for multiSURF neighborhoods

Uncomment this and debug. I don't understand what the issue is (bam).

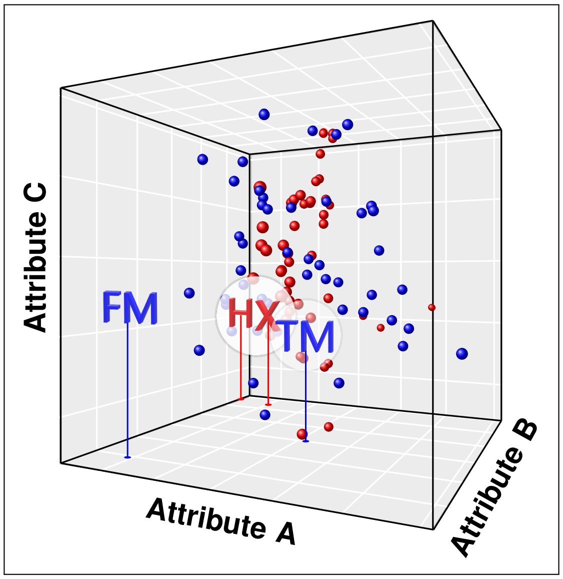

Fig 4. 3D AC view. Still working on this.

3 Derivation of means and standard deviations for metrics and data distributions

4 Optimal neighborhood parameters for detecting effects

k or α. Balancing blessing and curse of dimensionality.

5 ICA?