KD-Tree Algorithm for Propensity Score Matching

John R Hott, Nathan Brunelle, abhi shelat, Jeremy Rassen

April 16, 2012

Abstract

Objective. Propensity scores have been widely used in epidemiology to control for confounding in post-marketing Phase IV studies, however matching patients on multiple score components across multiple treatment groups becomes impractical. We present an algorithm that is expected-case quadratic on the number of participants per group.
Materials and Methods. Utilizing kd-tree data structures to provide efficient queries for nearby points and a search radius related to a best-guess match between participants in each treatment group, we reduce the number of participants that must be considered for each matching.
Results. Our algorithm outperforms brute force algorithms in the expected case, requiring only $O(n)$ space and $O(kdn^2)$ time compared with brute force's $O(n^{k+1})$ time, for k treatment groups. This difference is clearly seen in our empirical study of 1000 participants in 3 groups: our algorithm matches in 3.5 seconds compared to brute force's 19.53 hours.
Discussion. We prove the correctness of this approach, showing that each search radius considered contains the same match brute force choses for each participant.
Conclusion. While our approach does not match the performance of brute force in the worst case, we are no longer exponentially constrained by the number of groups in the expected case. Considering four groups of $5,000$ patients, that is a reduction from 625 trillion matches ($O(n^k)$) to 100 million ($O(n^2)$) and orders of magnitude shorter computation time.