Novel Approximation for Halfspace depth

Rasoul Shahsavarifar, David Bremner
Faculty of Computer Sceince, University of New Brunswick

Introduction

Data depth is a method to generalize the concept of rank in univariate data to multivariate data. By measuring the centrality of a data point with respect to a data set, data depth gives a center-outward ordering of points. Among different notions of data depth that have been defined over the last decades, we focus on **halfspace depth** and **Spherical depth** in this study**.**

In 1975, Tukey generalized the definition of univariate median and defined the halfspace median as a point in which the halfspace depth is maximized, where the halfspace depth is a multivariate measure of centrality of data points. In general, the halfspace depth of a query point q with respect to a given data set S is the smallest portion of data points that are separated by a closed halfspace through q.

In 2006, Elmore et al. defined another notion of data depth named spherical depth. The spherical depth of a query point q with respect to a data set S is defined as the probability that q is contained in a closed random hyperball with the diameter $x_i x_j$, where x_i and x_j are every two points in S.

The results of this study, can be applied to develop an **intrusion detection system.**

Definition

For a d-dimensional point q and a data set S $=\{x_1, \dots, x_n\} \subseteq \mathbb{R}^d$:

- The halfspace depth of q with respect to S is defined as:

$$HD(q;S) = \frac{1}{\lfloor n/2 \rfloor} \min \{|S \cap H|; H \in \mathbb{H}, q \in H\},$$

where $\mathbb{H}$ is the class of all closed halfspaces in $\mathbb{R}^d$.

- The spherical depth of q with respect to S is defined as:

$$SphD(q;S) = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} I(q \in Sph(x_i, x_j)),$$

where $Sph(x_i, x_j)$ is a hyperball with diameter $x_i x_j$.

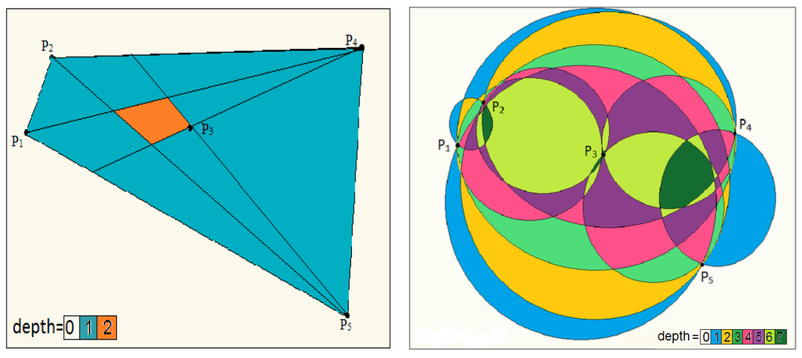


Fig: Halfspace depth (left) and spherical depth (right) of points in the plane with respect to $S = \{p_1, p_2, p_3, p_4, p_5\}$.

Computation

Over the last years, many different algorithms have been developed to compute the halfspace depth of a point in lower dimensions, or to find a point with the maximum value of halfspace depth. However, computing the halfspace depth in higher dimensions is a NP-hard problem. On the other hand, computing the spherical depth in dimension d takes only $O(dn^2)$.

Definition

For two nxn matrices P and Q, we define hamming distance d_H to measure the dissimilarity between these two matrices as follows:

$$d_H(P,Q) = \sum_{i,j} |P_{ij} - Q_{ij}|$$

Approximation

Due to the hardness of computing the halfspace depth problem, approximating this depth function is always of interests even in lower dimensions. In our research we propose a novel method to approximate the halfspace depth using another depth function (i.e. spherical depth). We have chosen the spherical depth because it is easy to implement even in higher dimensions. We approximate halfspace depth using spherical depth in two different ways.

❑ **Approximation using Euclidean distance**

In this method, we basically take the advantages of machine learning techniques. By training the approximation function and considering different criterions, we select the best model to approximate halfspace depth.

Halfspace depth=f(spherical depth),

where f is the approximation function. We use some test sets to compute the error of approximation.

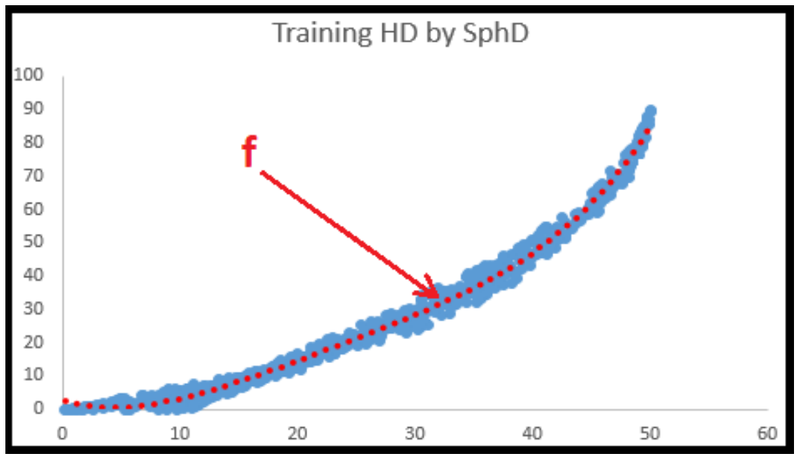


Fig: Function f approximates HD values using SphD values

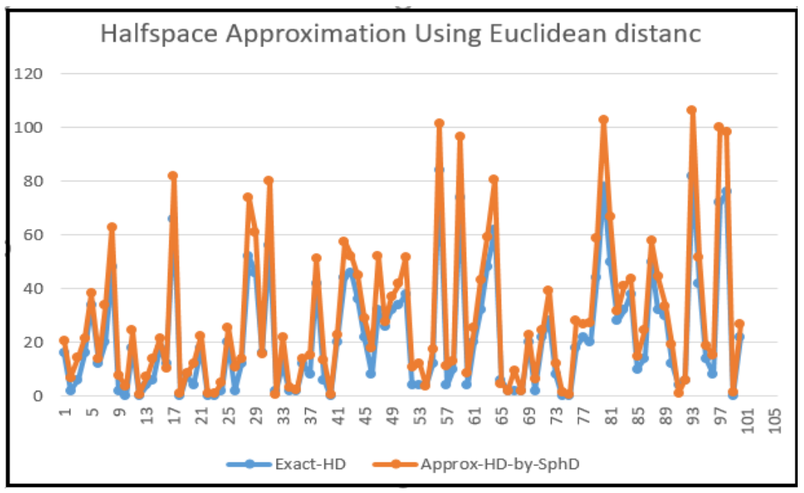


Fig: Comparing the approximated values vs exact values of HD

❑**Approximation using Hamming distance**

In this method, we focus on measuring the dissimilarity between the obtained **Posets** after applying the depth functions on a same data set. For every depth function D and a data set $S = \{x_1, .., x_n\}$, we define a matrix M_{nxn} as follows:

$$M[i][j] = \begin{cases} -1 & ; \ D(x_i;S) < D(x_j;S) \\ 1 & ; \ D(x_i;S) > D(x_j;S) \\ 0 & ; \ D(x_i;S) = D(x_j;S) \end{cases}$$

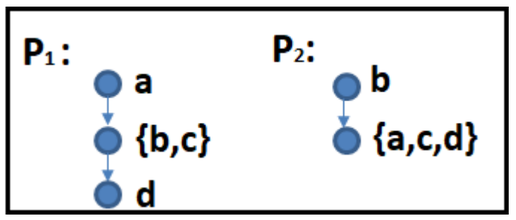
Using this representation, the hamming distance between depth functions D_1 and D_2 on the same data set S can be obtained by:

$$d_H(D_1, D_2) = \frac{1}{n^2 - n} \sum_{i,j} |P_{ij} - Q_{ij}|,$$

w here $n^2 - n$ is the normalization factor, P and Q are the matrices of D_1 and D_2, respectively.

Example

Suppose that P_1 and P_2 are two Posets obtained from applying D_1 and D_2 on data set $S = \{a, b, c, d\}$, respectively.



Using the matrix representations of P_1 and P_2, it can be figured out that the hamming distance $d_H(D_1, D_2) = 1/3$. The higher value of hamming distance means the less similarity in ranking behaviour of depth functions.

Conclusion and Ongoing Research

Based on the definitions, every depth function has a different behaviour in ranking data points in a given data set. We apply two different types of measures to approximate both the ranking behaviour and depth values of a data depth function. The concept of Euclidean distance helps us to approximate the halfspace depth values using the spherical depth values. On the other hand, we apply the hamming distance to approximate the ranking similarity between halfspace depth and spherical depth.

As an application of our results, we use these methods to approximate the **outliers** in a data set. This application motivates us to develop an **intrusion detection system**.

Reference

- Bremner, D. and Shahsavarifar, R., 2017. On the Planar Spherical Depth and Lens depth. *CCCG2017*.

- Bremner, D. and Shahsavarifar, R., 2017. An Optimal Algorithm for Computing the Spherical Depth of Points in the Plane. *arXiv preprint arXiv:1702.07399*.

Contact Info

Email: Ra.Shahsavari@unb.ca, Bremner@unb.ca

Faculty of Computer Science, University of New Brunswick, Fredericton, NB, Canada