The Hubness Phenomenon in High Dimensional Spaces

Name of First Author and Name of Second Author

Abstract Each chapter should be preceded by an abstract (10–15 lines long) that summarizes the content. The abstract will appear *online* at www.SpringerLink.com and be available with unrestricted access. This allows unregistered users to read the abstract as a teaser for the complete chapter. As a general rule the abstracts will not appear in the printed version of your book unless it is the style of your particular book or that of the series to which your book belongs.

Please use the 'starred' version of the new Springer `abstract` command for typesetting the text of the online abstracts (cf. source file of this chapter template `abstract`) and include them with the source files of your manuscript. Use the plain `abstract` command if the abstract is also to appear in the printed version of the book.

1 Introduction

[Describe the phenomenon, recent observations, and open questions we are investigating here]

Name of First Author
Name, Address of Institute, e-mail: name@email.address

Name of Second Author
Name, Address of Institute e-mail: name@email.address

2 Background and Related Work

3 Intrinsic Dimensionality via Hubness

3.1 Skewness vs. Feature Ranking - What is the Right Way to Rank Features?

[Local Skewness plots, etc.]

3.2 Supervised vs. Unsupervised Methods

[Compare methods for feature relevance, cluster properties in original and subspace, etc.]

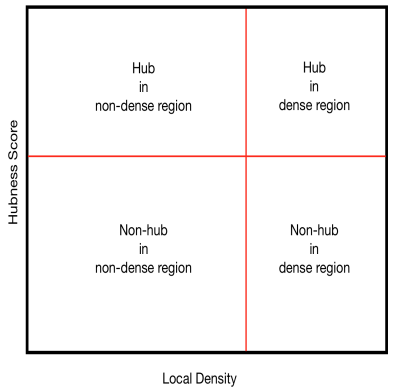
4 Hubs, Density, and Clustering

To investigate the relationship between density hubness scores, and clustering, data sets of 30, 60, and 100 relevant dimensions were created. For each of these dimensions, two types of data sets were generated: (1) two Gaussian spheres separated by 10 units in the first coordinate and (2) two uniform cubes separated by one unit along its first dimension. In order to emulate varying densities, one cluster was designed to contain 1000 points and the second varied in size (2000, 3000, 4000, and 5000). There was a resulting total of 24 data sets given the different dimensions, distribution of points, and varying densities. For the rest of the chapter, the Gaussian data sets will be denoted as $G_{d,N}$ and the uniform ones as $U_{d,N}$, where d is the dimension and N is the number of points. In general, any d-dimensional data of N points will be denoted as $X_{d,N}$.

4.1 Hubness and Data Density

The first question to be addressed is whether hubness scores and density are the same measure. One way to see this is by plotting density against hubness scores. Suppose that there is a low density area, high density area, a low hubness score area and a high hubness score area, shown in Fig. 4.1. Intuition dictates that if density and hubness score were the same, all the points will lie within the first (high density and high hubness scores) and third (low density and low hubness scores) quadrants. If density and hubness score were the opposite, all the points will lie within the second

(low density and high hubness scores) and fourth (high density and low hubness scores) quadrants. However, if the points are all over the four quadrants, then it might be telling of a different relationship between density and hubness scores.



For the experiments, the hubness scores N_k were calculated for all the points with $k = 5, 10$, and 50. As far as the density estimation, $q_k(x) = \frac{1}{||x-x_k||}$, where x is a data point and x_k is the kth nearest neighbor of x, was chosen for its efficiency and since more sophisticated non-parametric consistent density estimators also suffer from the curse of dimensionality. All the density estimation plots used q_{16}. To determine what is "high density" and "high hubness score," a threshold value of two standard deviations above the mean was calculated. A global threshold value, meaning that N_k and q_{16} were calculated for all the points regardless of class membership and then averages and standard deviations were calculated, can be seen in Fig.1.

[Distance plots, Correlation of hubs to density, etc.]

4.2 Hubness and Clusters

[How do hubs relate to special points from various clustering methods (DBSCAN core points, outliers, etc.)?]

4.3 Other Experiments PlaceHolder

[These experiments could go into previous subsections as deemed fit:Landscape of hubs for different structured data, purity of reverse neighbors]

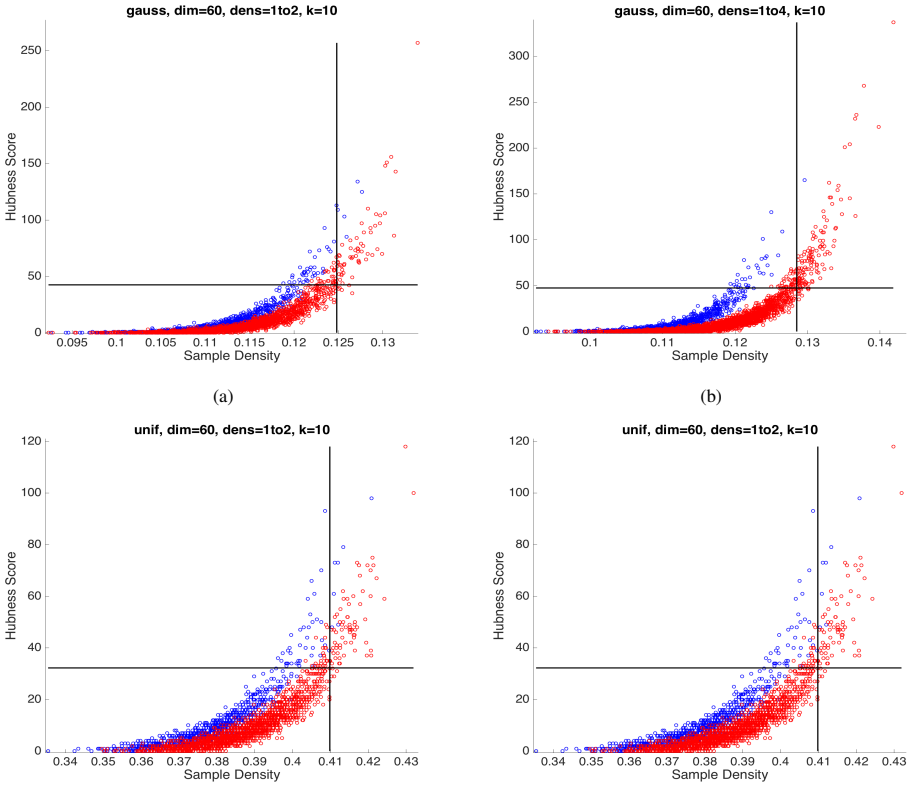


Fig. 1 Density, q_{16}, plotted against hubness score, N_{10}, for data sets (a) $G_{60,3000}$, (b) $G_{60,5000}$, (c) $U_{60,3000}$, and (d) $U_{60,5000}$.

5 Discussion

6 Conclusion and Future Work