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, as shown in Fig. 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.

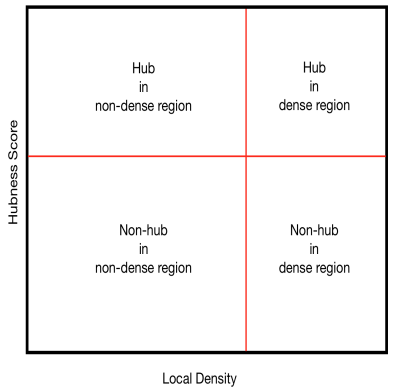

Fig. 1 An illustration of the four quadrants, which are formed by density and hubness scores thresholds, that will be seen in the following experiments where density will be plotted against hubness scores.

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. Taking into account the different types of data, densities, dimensions and k's, there were 72 total plots, and Fig.2 shows representative plots of all the experiments conducted. In these figures, global thresholds for density and hubness scores were plotted in black. Global thresholds refer to the fact that N_k and q_{16} were calculated for all the points regardless of class membership and then averages and standard deviations were calculated. The blue and red represent the class membership of each point, which is known from how the data was created.

The most obvious aspects of the data plotted in Fig. 2 is the strong positive correlation and the "wings" formed by the two clusters. The positive correlation between these two measures seem to be close to quadratic. Also, the experiments show that the bigger the difference in densities between the two clusters, the more of a gap there is, and hence forming the "wings." Due to the curse of dimensionality, these wings are more common in the 30 and 60 dimensional data than they are in the 100 dimensional data.

The next observation has to do with the four quadrants of Fig. 1, which can be seen in each plot of Fig. 2. In all 72 experiments, more than 90% of the data lies in the third quadrant (non-hub and low density), which is due to how we defined the

thresholds. The interesting observations come from the proportions of data in the other quadrants. In 23 out of the 36 Gaussian experiments, i.e. 63.89% of the experiments, a larger proportion of hubs in the high density region. This suggests that for the Gaussian distributions, hubs are more likely to be in a high density region. For the Uniform data, 30 out of 36 experiments, or about 83% of the experiments, had a larger proportion of hubs in the low density region, which means that hubs are most likely in low density regions. This would suggest that hubness scores and density are not exactly the same measure.

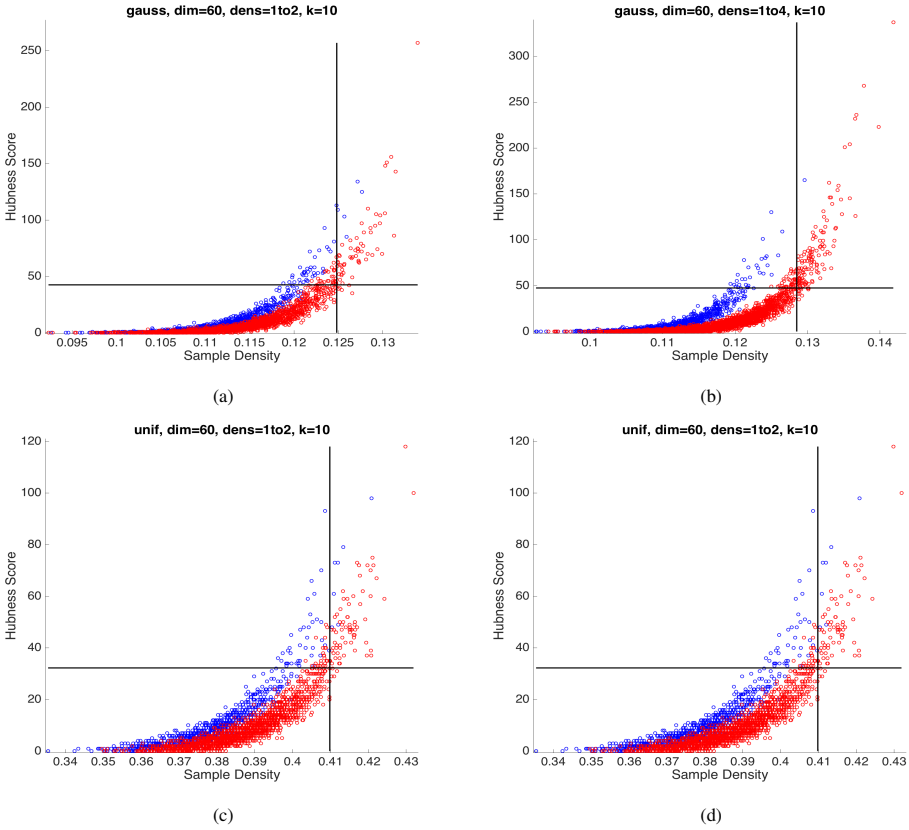

Fig. 2 Density, q_{16}, plotted against hubness score, N_{10}, for data sets (a) $G_{60,3000}$, (b) $G_{60,5000}$, (c) $U_{60,3000}$, and (d) $U_{60,5000}$. The true class is depicted in the two colors of the points and the global thresholds are plotted in black.

The problem with just looking at a global threshold is that it does not account for the difference in densities between the two clusters. Hence, the same experiments were performed but this time with local thresholds, meaning that the same N_k and q_{16} were used, but this time the mean and standard deviation were calculated for each class. The thresholds for the light blue cluster were plotted in dark blue and the

ones for the pink one were plotted in red. Some of the resulting plots can be seen in Fig 3. For these experiments, the Gaussian experiments had similar results as the global ones. About 64% of the Gaussian experiments had most of its hubs in the high density region. However, while the Uniform experiments with "global hubs," i.e. the global threshold, were mainly in the global non-dense regions, for about 85% of the "local" experiments had most of the local hubs in the local ?dense? region. This again supports that hubness scores and density are closely related.

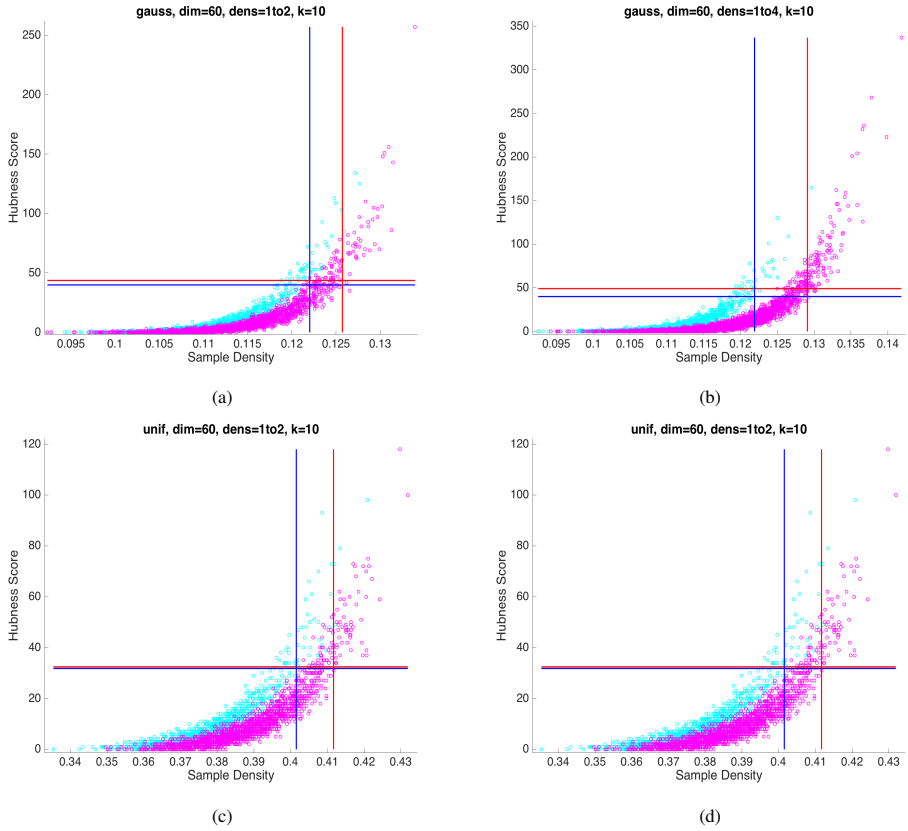

Fig. 3 Density, q_{16}, plotted against hubness score, N_{10}, for data sets (a) $G_{60,3000}$, (b) $G_{60,5000}$, (c) $U_{60,3000}$, and (d) $U_{60,5000}$. The true class is depicted in the two colors: pink and blue. Class thresholds are plotted in red (for the pink class) and blue (for the blue class).

[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]

5 Discussion

6 Conclusion and Future Work