Density and Hubness

Libby Beer (libby.beer@gmail.com)

Jesse Metcalf-Burton (datasciencejess@gmail.com)

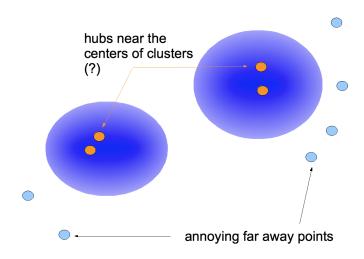
Marilyn Vazquez ()

July 21, 2017

1 / 10

Distances between points? Hubs?

Can hubs help us cluster?



Distances with synthetic data

n-hubness//users/jesse//graphs/gaussians_1000_5000_60.mat.10.png Distances between (all) points IntrA-cluster distances 0.20 4.0 Density 9:0 0.2 80 00 10 20 10 15 20 Distance between points Distance between same-cluster hubs Distances between hubs IntER-cluster distances 9. 9.5 4.0 4.0 0.3 6.0 Density 0.2 0.2 5 5 8 0.0 15 5 15 5 10 20 10 20

Distance between diff-cluster hubs Hubness

Density/Clustering ()

Distance between hubs

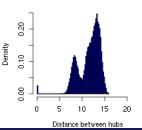
July 21, 2017

Distances with (other) synthetic data

ubness//users/jesse//graphs/synthetic_data_30_50_overlap30.mat.10.png Distances between (all) points IntrA-cluster distances 8 0.20 Density 0.10 0.10 80 900 10 15 20 n 10 20

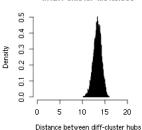


Distance between points



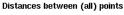
IntER-cluster distances

Distance between same-cluster hubs



Distances with real data

:7/wisdm-hubness//users/jesse//graphs/data_spam.mat.10.png

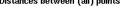


2.0 IŲ.

9

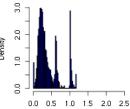
9 8

> 0.0 0.5 1.0 1.5 2.0





IntrA-cluster distances

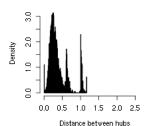


Distance between same-cluster hubs

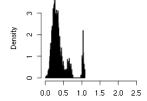
IntER-cluster distances

Distances between hubs

Distance between points

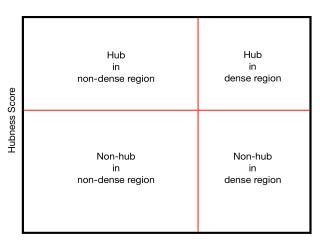


Distance between diff-cluster hubs Hubness



Do hubs appear in high density regions?

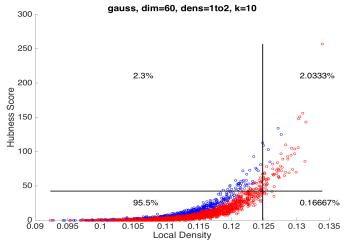
Intuition



Local Density

Do hubs appear in high density regions?

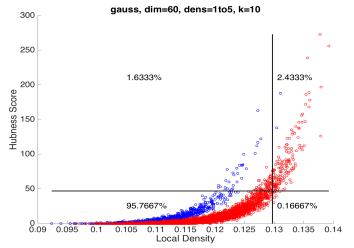




Notice: (1) Proportion of hubs in the two density regions, (2) proportion of high density points that are hubs, and (3) more hubs in the denser cluster.

Do hubs appear in high density regions?

Answer: Sometimes!



Notice: (1) A LOT more hubs in the denser cluster and (2) the rise of two "wings" that correspond to the two different cluster.

Observations

From the experiments on the Gaussian data, we observe that:

- About 13 out of 36 experiments, there was a larger proportion of hubs in the low density region i.e. hubness scores and density are not exactly the same thing
- About 23 out of 36 experiments, there was a larger proportion of hubs in the high density region i.e. if a point is a hub, it is more likely to be in a high density region.
- All our experiments show that high density regions are mostly composed of hubs i.e. if a point is in a high density region, it is most likely a hub.

Density/Clustering ()

Future Work

- More experiments of local density versus hubness scores for other density types e.g. uniform density
- Verify if patterns also show up in real data
- Data with more clusters to see if "wings" arise for each cluster
- Test model selection via clustering only hubs
- Comparison of hubs to core points, outliers, etc.

Density/Clustering ()