Improving Recommender systems by reducing Hubness

Donthu Vamsi Krishna (15111016), Dhekane Eeshan Gunesh (13248) Advisor: Prof. Piyush Rai
Indian Institute of Technology Kanpur

Introduction

- Hubs are the data points which occur in the nearest neighbours of most of the data points though the data points are not similar making the nearest neighbour relations asymmetric.
- Hubness problem mostly occurs while dealing with high dimensional data.
- Since most of the recommender systems deal with high dimensional data and rely on nearest neighbour relations, their performance degrades due to the presence of hubs.

Previous Work

Study of effects of hubs in various domains.

- [Acoutier, Pachet] have observed certain songs that were similar to a large number of other songs with respect to the used audio-similarity functions.

Connections with previous works.

- [Pampalk, Karydis and Flexer] showed that hubs can be viewed as False Positives when are considered in context of classification problem.
- [Berenzwig] has commented on possible relation between hub problems and the high dimensionality of the feature space.

Dataset

We have used the MovieLens data set.

- We have used 3 datasets: 100k, 1m and 10m.
- We also took subset of 10k ratings from 100k dataset.

We have also used IRIS flower dataset

- We created synthetic dataset of 15k features using IRIS dataset containing 150 features.
- This synthetic dataset is created by adding small variation to the original dataset and maintaining its corresponding flower class.

Measuring Hubness

Skewness Measure

$$N_k(x) = \sum_{i=1}^{n} I_{k,i}(x) \quad \text{Where} \quad I_{k,i}(x) = \begin{cases} 1 & \text{If } x \text{ appears in the } k \text{ nearest neighbors list of } x_i \\ 0 & \text{otherwise} \end{cases}$$

Goodman-Kruskal Index

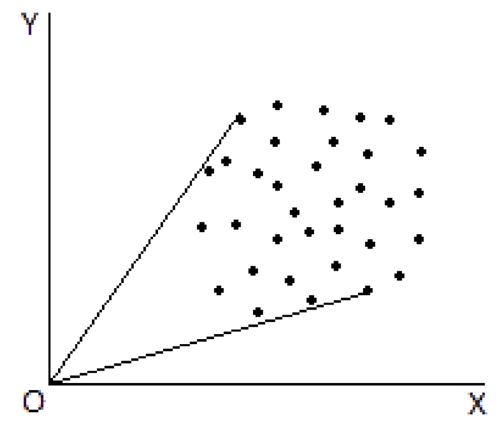
$$I_{Gk} = \frac{N_C - N_D}{N_C + N_D}$$

Where N_C is the number of concordant tuples,
N_D is the number of discordant tuples.

A tuple $T=(i,j,k,l)$ chosen such that x_i, x_j belong to same class and x_k, x_l belong to different classes.
T is called as concordant tuple iff $d(x_i, x_j) < d(x_k, x_l)$ otherwise T is called as discordant tuple

Reducing Hubness

Centering & Weighted Centering



$$x^{\text{cent}} = x - \bar{x}$$

$$x^{\text{weighted}} = x - \bar{x}^{\text{weighted}}$$

$$\bar{x}^{\text{weighted}} = \sum_{i=1}^{n} w_i x_i$$

$$w_i = \frac{d_i^{\gamma}}{\sum_{j=1}^{n} d_j^{\gamma}}$$

$$d_i = \sum_{j=1}^{n} \langle x_i, x_j \rangle = n \left\langle x_i, \frac{1}{n} \sum_{j=1}^{n} x_j \right\rangle$$

Local Scaling & Global Scaling

Make the nearest neighbor relations close to symmetric.

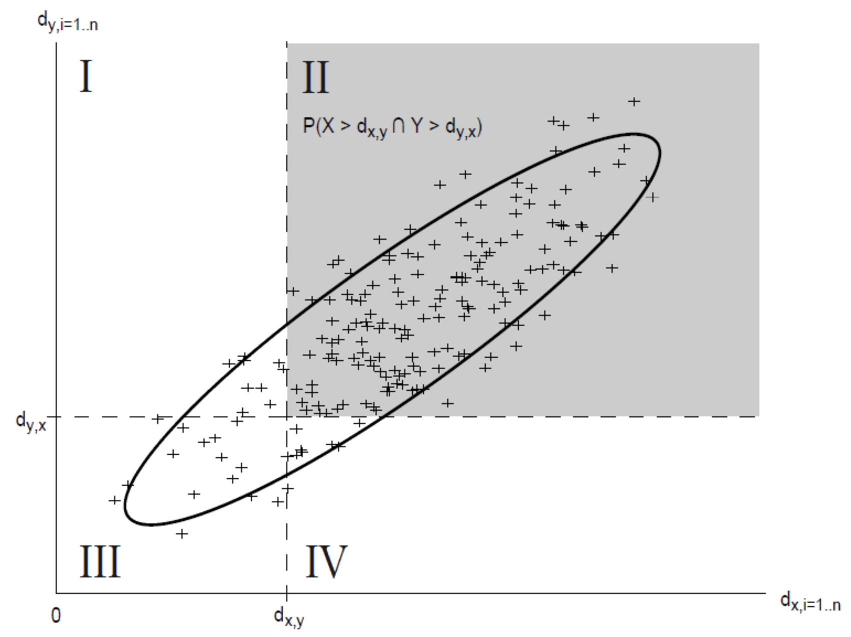
$$LS(d_{x,y}) = e^{-\frac{d_{x,y}}{\sigma_x}\frac{d_{x,y}}{\sigma_y}}$$

Where σ_z is the standard deviation in the distance of point z from its k nearest neighbors.

$$MP(d_{x,y}) = P(X > d_{x,y} \text{ and } Y > d_{y,x})$$

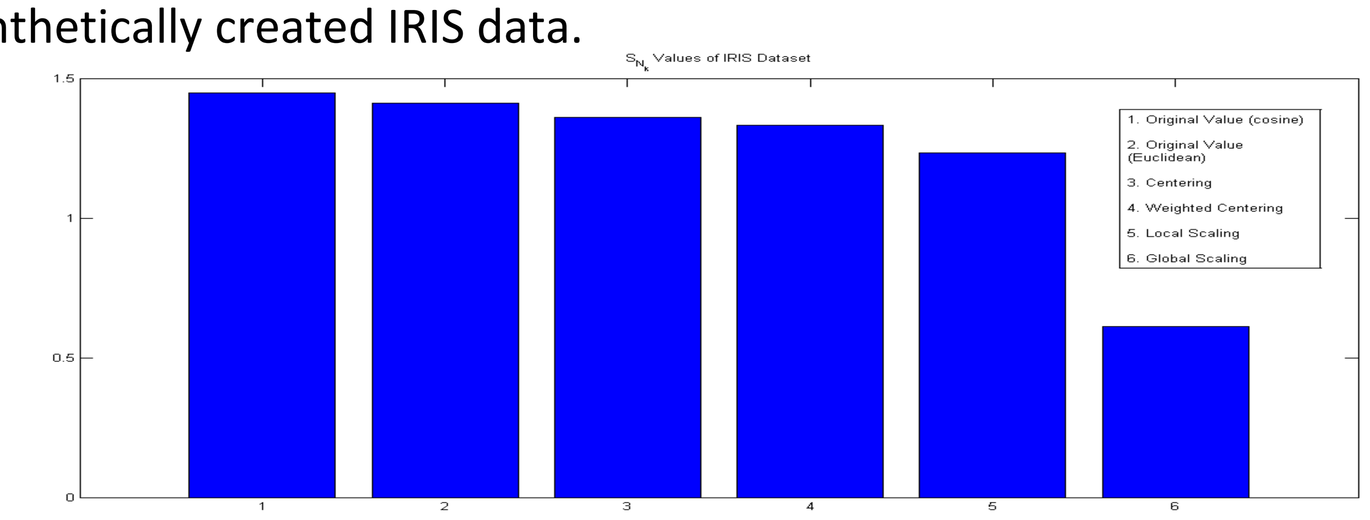
Assuming Independence, we get

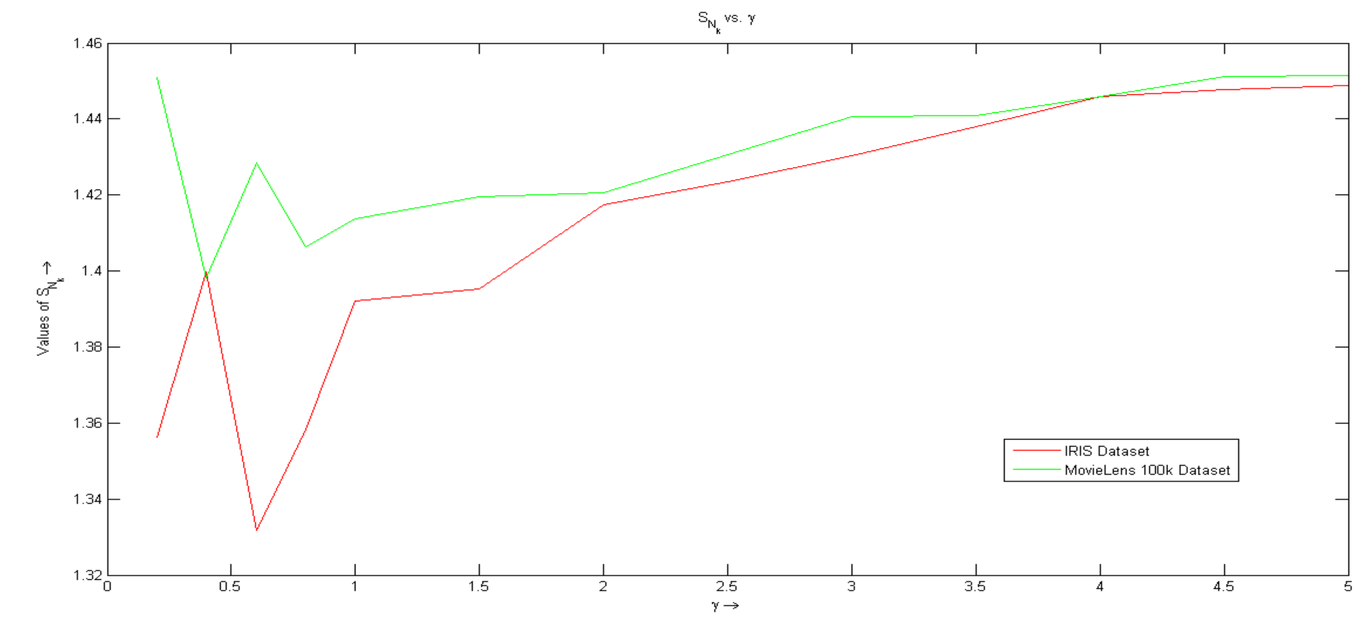
$$MP(d_{x,y}) = P(X > d_{x,y}).P(Y > d_{y,x})$$



Experiments & Results

We ran our experiments on 10k movie lens subset data and synthetically created IRIS data.





Conclusions

The best value of γ in weighted centering lies between 0 and 1 and can be chosen by preferring local minima.
Local and global scaling reduces the amount of hubness to a great extent, but require high computational capabilities.
The methods used to measure hubness is to be chosen on the basis of our requirements.

References

[1] Suzuki, Ikumi, et al. "Centering Similarity Measures to Reduce Hubs." EMNLP. Vol. 13. 2013.

[2] Radovanović, Miloš, Alexandros Nanopoulos, and Mirjana Ivanović. "Hubs in space: Popular nearest neighbors in high-dimensional data." The Journal of Machine Learning Research 11 (2010): 2487-2531.

[3] Schnitzer, Dominik, et al. "Local and global scaling reduce hubs in space." The Journal of Machine Learning Research 13.1 (2012): 2871-2902.

[4] Knees, Peter, Dominik Schnitzer, and Arthur Flexer. "Improving neighborhood-based collaborative filtering by reducing hubness." Proceedings of International Conference on Multimedia Retrieval. ACM, 2014.