Reducing Hubness in High Dimensional Data Spaces

Anand Singh Kunwar, Neeraj Kumar

Under the Guidance of Prof. Piyush Rai

Abstract

High-dimensional data arise naturally in many domains, and have regularly presented a great challenge in recommendation systems due to the presence of hubs. Hubness has recently been identified as a general problem of high dimensional data spaces, manifesting itself in the emergence of objects, so-called hubs, which tend to be among the k nearest neighbours of a large number of data items[1]. In this project, we propose to reduce hubness in high dimensional data spaces using different techniques and testing them on some available data.

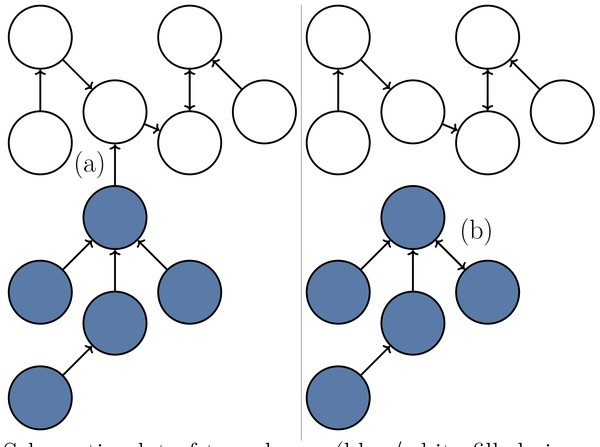
Introduction

In some of the recent publications, the hubness phenomena has been described as a general problem of machine learning in high-dimensional data space. Hubs are data point which keep appearing frequently in the nearest neighbour of many other points. As a result, in recommender systems hub points are recommended again and again and some points (anti hubs) are never recommended. Hubness is related to the phenomenon of concentration of distances, which is the fact that all points are at almost the same distance to each other for dimensionality approaching infinity (Francois et al., 2007) [2].

Related Work

Scaling methods have been used in some papers to reduce hubness in music recommendation system. Shared Nearest Neighbours approach was first used to symmetrize the nearest neighbour relations. Local Scaling methods by Zelnic et al., where the distances are scaled according to the neighbourhood information of the data point have been employed in the past. The rescaling was usually done by dividing the distance with the distance of the k^{th} nearest neighbour's distance. Global Scaling methods employed transformation of the points into a probability that two points are in the nearest neighbourhood of each other.

Demonstration



Schematic plot of two classes (blue/white filled circles). Each circle has its nearest neighbor marked with an arrow: (a) violates the pairwise stability clustering assumption, (b) fulfills the assumption. In many classification and retrieval scenarios, (b) would be the desired nearest neighbor relation for the data set.

Methods

- **Scaled Shared Nearest Neighbors**

$$SS_2(i,j) = d(i,j).SS(i,j) \tag{1}$$

This tweak on $SS(i,j)$ combines the scaling with shared nearest neighbor heuristic.

- **Local Scaling**

We scale each distance $d(s_i, s_j)$ with scaling factor σ_i, so new distance is $d(s_i, s_j)/\sigma_i$

$$\sigma_i = d(s_i, s_K) \tag{2}$$

where s_K is the K^{th} nearest neighbour of s_i

- **Global Scaling**

We transform the distance between points x and y into the probability that y is the closest neighbour to x

$$MP(d_{x,y}) = P(X > d_{x,y} \cap Y > d_{x,y}) \tag{3}$$

We assume independence of distributions of $P(X)$ and $P(Y)$, $\therefore$ our new Mutual Proximity is defined as below:

$$MP_2(d_{x,y}) = P(X > d_{x,y}).P(Y > d_{y,x}) \tag{4}$$

Hubness Measure

- **Skewness Measure** (S^k)

We define N_k (K-occurence) = Number of times x occurs in the k-NN lists of all other data points. We define Hubness S^k using the 2010 Radovanovic et al. [3]:

$$S^k = \frac{E[(N_k - \mu_{N_k})^3]}{\sigma_{N_k}^3} \tag{5}$$

- **Goodman-Kruskal Index** (I_{GK})

We use the set of concordant points Q_c and discordant points Q_d:

$$I_{GK} = \frac{|Q_c| - |Q_d|}{|Q_c| + |Q_d|} \tag{6}$$

Results

We used Arcene Dataset, Urban Land Cover Dataset

Arcene Dataset

- Euclidean Distance Metric

Value of $k = 5$

Method	Hubness	I_{GK}	Accuracy
Original	0.25	0.067	75.00%
SNN	0.36	0.058	61.00%
SSNN	0.16	0.073	66.00%
LS	-0.01	0.064	79.00%
MP	0.01	0.096	80.00%

- Manhattan Distance Metric

Value of $k = 5$

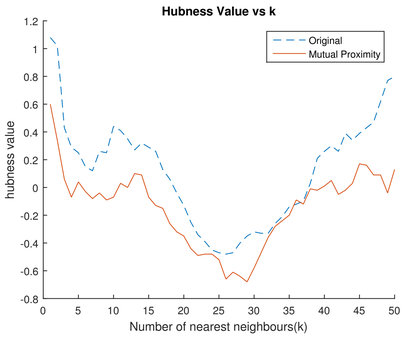
Method	Hubness	I_{GK}	Accuracy
Original	0.46	0.056	78.00%
SNN	0.36	0.058	61.00%
SSNN	0.20	0.068	70.00%
LS	0.07	0.066	80.00%
MP	0.44	0.045	69.00%

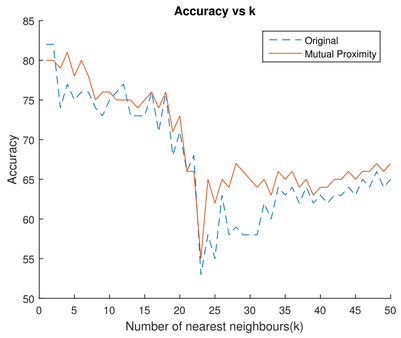
Urban Land Cover Dataset

- Euclidean Distance Metric

Value of $k = 5$

Method	Hubness	I_{GK}	Accuracy
Original	0.37	0.102	44.58%
SNN	0.38	0.079	37.48%
LS	-0.10	0.171	46.35%
MP	-0.15	0.138	45.17%





Conclusion

In above experiments with public machine learning databases we have shown that both local and global scaling methods lead to: (i) a significant decrease of hubness, (ii) an increase of k-nearest neighbor classification accuracy, and (iii) a strengthening of the pairwise class stability of the nearest neighbors.

References

[1] Dominik Schnitzer, Arthur Flexer, Markus Schedl, and Gerhard Widmer.
Local and global scaling reduce hubs in space.
The Journal of Machine Learning Research, 13(1):2871–2902, 2012.

[2] Damien Francois, Vincent Wertz, and Michel Verleysen.
The concentration of fractional distances.
Knowledge and Data Engineering, IEEE Transactions on, 19(7):873–886, 2007.

[3] Miloš Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic.
Hubness in the context of feature selection and generation.
In *Proceedings of the SIGIR Feature Generation and Selection for Information Retrieval Workshop (FGSIR)*, page 9, 2010.