Reducing Hubness in High Dimensional Data Spaces

Anand Singh Kunwar 13110
Neeraj Kumar 13427

Under 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[6]. In this project, we propose to reduce hubness in high dimensional data spaces using different techniques and testing them on some available data.

Contents

1 Motivation **2**

2 Introduction **2**
2.1 Demonstration of the Problem 3

3 Related Work **3**

4 Methodology **3**
4.1 Shared Nearest Neighbours 4
4.2 Local Scaling 4
4.3 Global Scaling - Mutual Proximity 4
4.4 Hubness Measure 5
4.5 Goodman-Kruskal Index 5

5 Results **6**
5.1 Arcene Dataset 6
5.1.1 Euclidean Distance Metric 6
5.1.2 Manhattan Distance Metric 7
5.1.3 Chebyshev Distance Metric 7
5.1.4 Diagonal Grid Distance Metric 8
5.2 Urban Land Cover Dataset 8
5.2.1 Euclidean Distance Metric 8
5.2.2 Manhattan Distance Metric 8
5.2.3 Chebyshev Distance Metric 8
5.2.4 Diagonal Grid Distance Metric 9

6 Conclusion **9**

1 Motivation

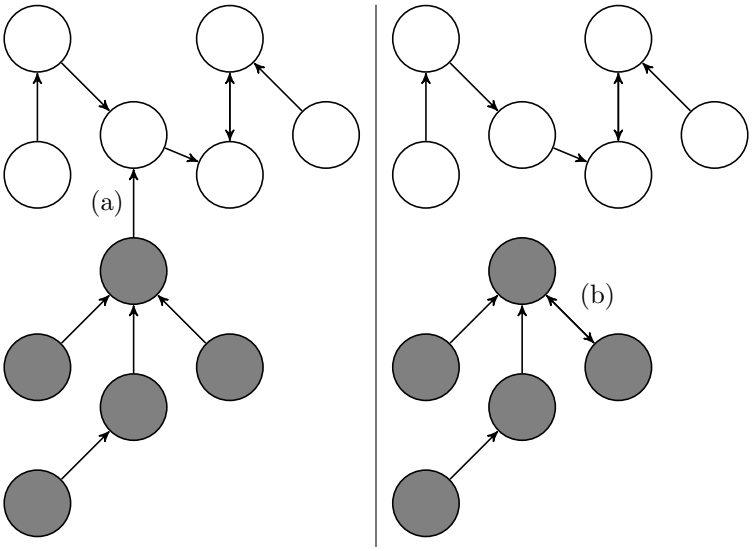
In a 2010 paper, Radovanovic et al.[5] described hubness as a property of data vectors to lie in unexpectedly in many k-nearest neighbours list of other points in data. This causes skewness in the nearest neighbour lists where some points lie in the nearest neighbours of many points (hubs) whereas some points don't lie in any of the nearest neighbours of data points (anti-hubness). So our aim is to reduce hubness in data points by modifying our distance metric using various methods.

2 Introduction

In some of the recent publications, the so called hubness phenomena has been described as a general problem of machine learning in highdimensional 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) [1].

Radovanovic et al. (2010)[5] presented that some points are expected to be closer to the center of all data points and these points are candidates of hub points while some points which are away from centre of data points are preferable for anti hub points. Arthur Flexer in his paper has stated that hubness phenomena is because of extrinsic dimensionality of dataset. The extrinsic dimension is the actual number of dimensions of a data space, the intrinsic dimension is the, often much smaller, number of degrees of freedom of the submanifold in which the data space can be represented (Francois et al., 2007)[1]. In our implementation, we have also measured the extrinsic dimension of dataset and analysed the result. As already said, we can reduce the effect of hubness if we symmetrize the nearest neighbour relations. For this we have used different methods like Shared Nearest Neigbour, Local Scaling, Mututal Proximity. As the name suggests, the shared near neighbor (SNN) similarity is based on computing the overlap between the k nearest neighbors of two objects. Local scaling tries to symmetrize the nearest neigbour relationship using local neighborhood information to rescale distances between data points. Another variant of above method is Mutual Proximity, which is a global scaling method and uses probability distribution of distances between data points. These methods are unsuperwised methods because they dont use class labels and uses only similarity between data points and scale them.

2.1 Demonstration of the Problem



Schematic plot of two classes (black/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.

3 Related Work

Scaling methods have been used in some papers to reduce hubness in music recommendation system[4]. Shared Nearest Neighbours approach was first used to symmetrize the nearest neighbour relations. Local Scaling methods by Zelnic et al.[7], 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. Schnitzer et al.[6] defined the Mutual Proximity of data points which is shown further in the report.

4 Methodology

We have extended it further and in addition to implementation of above techniques we have also tweaked these methods, have used different distance measures like Euclidean, Manhattan, Chebyshev and Grid Diagonal distance metric and have analysed those results. We have used different datasets and these

methods have not only reduced the hubness but also have improved knearest neighbour classification accuracy in most of the instances.

4.1 Shared Nearest Neighbours

The first thought to remove asymmetric nearest neighbour relations is to use something like 'shared nearest neighbour' approaches. The paper on SNN (shared nearest neighbours) by Jarvis et al.[3] suggested computation of overlapping k nearest neighbours of two data points. For data points i, j we calculate K_i, K_j where K_i corresponds to the set of k nearest neighbours of i. We define the new $SS(i, i)$ as the following. We also tweaked this to form SS_2 which is also stated below:

$$\begin{aligned} T_{i,j} &= K_i \cap K_j \\ SS(i,j) &= 1 - \frac{|T_{i,j}|}{k} \end{aligned} \tag{1}$$

$$SS_2(i,j) = d(i,j).SS(i,j) \tag{2}$$

This tweak on $SS(i, j)$ combines the scaling with shared nearest neighbor heuristic. The data shows it to improve accuracy and often reduce hubness.

4.2 Local Scaling

We then move on to local scaling methods, which compute scaled distances using the neighbourhood information of the data point. Zelnik et al.[7] proposed a local scaling parameter σ_i and rescaled the distance $d(s_i, s_j)$ of data points s_i, s_j to $d(s_i, s_j)/\sigma_i$. The choice of σ_i can be $d(s_i, s_K)$ where s_K is the K^{th} nearest neighbour of point s_i

$$\sigma_i = d(s_i, s_K) \tag{3}$$

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

In some cases, σ_i is also defined as the average distance of the K-nearest neighbours of data point s_i

$$\sigma_i = \frac{\sum_{j=1}^{K} d(s_i, s_j)}{K} \tag{4}$$

4.3 Global Scaling - Mutual Proximity

In global scaling, we transform the distance between points x and y into the probability that y is the closest neighbour to x given the distribution of the distances of all points to x in the data set and then we combine these probabilistic distances from x to y and y to x via their joint probability. The result is a general unsupervised method to transform arbitrary distance matrices to matrices of probabilistic mutual proximity (MP) [6].

We now compute the probability that y is the nearest neighbor of x given $P(X)$ which is the probability distribution function defined by $d_{(x,i=1..n)}$ and

probability that x is the nearest neighbor of y given $P(Y)$ which is the probability distribution function defined by $d_{(y,i=1..n)}$. Now we compute the Mutual Proximity of $MP(d_{x,y})$

$$MP(d_{x,y}) = P(X > d_{x,y} \cap Y > d_{x,y}) \tag{5}$$

Since this requires estimation of a joint distribution of $P(X, Y)$ for all distance pairs $d_{x,y}$ and is computationally heavy. We assume independence of distributions of $P(X)$ and $P(Y)$, $\therefore$ our new Mutual Proximity is defined as below:

$$MP_2(d_{x,y}) = P(X > d_{x,y}).P(Y > d_{y,x}) \tag{6}$$

In the paper by Schnitzer et. all of 2012 [6], it is shown that the assumption of independence in the computation of Mutual Proximity does not affect the results by such a big factor.

4.4 Hubness Measure S^k

To define the Hubness measure we first define the K-occurrence of an object x. N_k (K-occurence) = Number of times x occurs in the k-nearest neighbour lists of all other data points. We define Hubness S^k using the 2010 Radovanovic et al. [5] paper which is the skewness of the k-occurrences N_k:

$$S^k = \frac{E[(N_k - \mu_{N_k})^3]}{\sigma^3_{N_k}} \tag{7}$$

The high positive values indicate high hubness.

4.5 Goodman-Kruskal Index I_{GK}

Goodman-Kruskal index [2] is a measure of quality of clusters formed for a distance matrix d.

We define concordant points Q_c and discordant points Q_d as follows:

- Tuple $(d_{i,j}, d_{k,l}) \in Q_c$ if items i and j are from the same class, items k and l are from different classes and $d_{i,j} < d_{k,l}$
- Tuple $(d_{i,j}, d_{k,l}) \in Q_d$ if items i and j are from the same class, items k and l are from different classes and $d_{i,j} > d_{k,l}$
- For all the rest of the cases the tuple is neither concordant nor disconcordant

We use the set of concordant points Q_c and discordant points Q_d to define the Goodman Kruskal Index as follows:

$$I_{GK} = \frac{|Q_c| - |Q_d|}{|Q_c| + |Q_d|} \tag{8}$$

5 Results

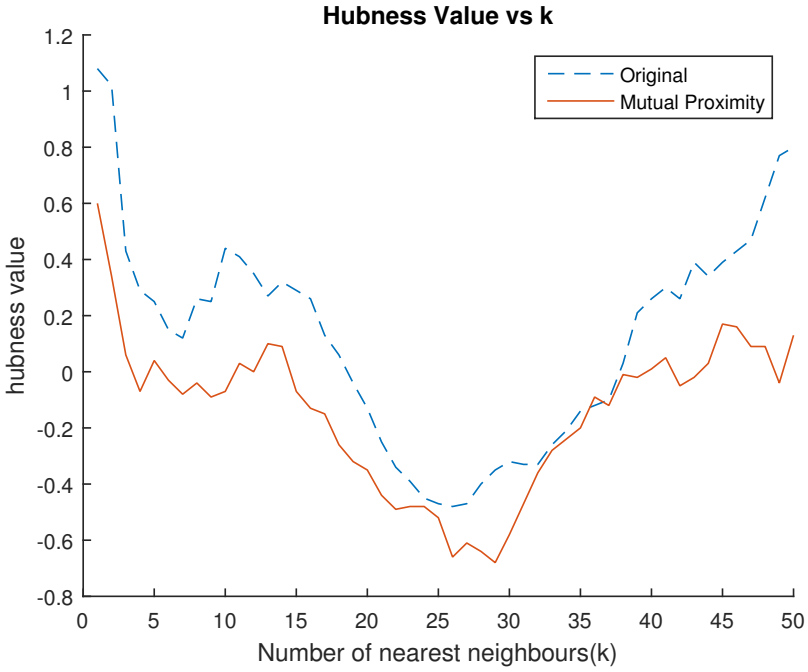
5.1 Arcene Dataset

This dataset is one of 5 datasets of the NIPS 2003 feature selection challenge. This data is from mass spectrometry which is used to differentiate between cancer and non-cancer patients. This classification problem has 2 classes - cancer and non-cancer patients.

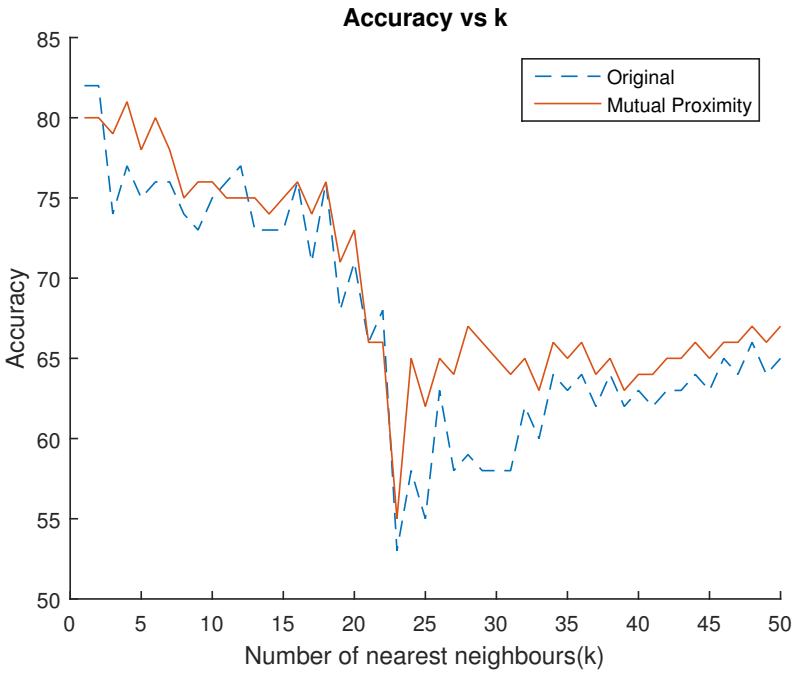
5.1.1 Euclidean Distance Metric

Value of $k = 5$

Method	Hubness	Goodman-Kruskal Index	Accuracy
Original	0.25	0.067	75.00%
Scaled SNN	0.16	0.073	66.00%
SNN	0.36	0.058	61.00%
Local Scaling	-0.01	0.064	79.00%
Mutual Proximity	0.01	0.096	80.00%



Hubness vs k for Euclidean Distance Metric (Arcene Data Set)



Accuracy vs k for Euclidean Distance Metric (Arcene Data Set)

5.1.2 Manhattan Distance Metric

Value of $k = 5$

Method	Hubness	Goodman-Kruskal Index	Accuracy
Original	0.46	0.056	78.00%
Scaled SNN	0.20	0.068	70.00%
SNN	0.36	0.058	61.00%
Local Scaling	0.07	0.066	80.00%
Mutual Proximity	0.44	0.045	69.00%

5.1.3 Chebyshev Distance Metric

Value of $k = 5$

Method	Hubness	Goodman-Kruskal Index	Accuracy
Original	1.16	0.045	67.00%
Scaled SNN	0.94	0.040	61.00%
SNN	0.79	0.036	57.00%
Local Scaling	0.59	0.051	66.00%
Mutual Proximity	0.43	0.047	64.00%

5.1.4 Diagonal Grid Distance Metric

Value of $k = 5$

Method	Hubness	Goodman-Kruskal Index	Accuracy
Original	1.13	0.045	66.00%
Scaled SNN	0.81	0.040	64.00%
SNN	0.63	0.038	60.00%
Local Scaling	0.59	0.051	65.00%
Mutual Proximity	0.44	0.045	65.00%

5.2 Urban Land Cover Dataset

This dataset contains data to classify a high resolution aerial image into 9 types of urban land. The land cover classes are: trees, grass, soil, concrete, asphalt, buildings, cars, pools, shadows. The information given includes multi-scale spectral, size, shape, and texture information.

5.2.1 Euclidean Distance Metric

Value of $k = 5$

Method	Hubness	Goodman-Kruskal Index	Accuracy
Original	0.37	0.102	44.58%
SNN	0.38	0.079	37.48%
Local Scaling	-0.10	0.171	46.35%
Mutual Proximity	-0.15	0.138	45.17%

5.2.2 Manhattan Distance Metric

Value of $k = 5$

Method	Hubness	Goodman-Kruskal Index	Accuracy
Original	0.53	0.208	58.97%
SNN	0.54	0.139	48.32%
Local Scaling	0.04	0.298	59.17%
Mutual Proximity	0.07	0.247	60.36%

5.2.3 Chebyshev Distance Metric

Value of $k = 5$

Method	Hubness	Goodman-Kruskal Index	Accuracy
Original	0.43	0.068	37.08%
SNN	0.28	0.075	32.35%
Local Scaling	-0.17	0.133	38.66%
Mutual Proximity	-0.16	0.105	38.07%

5.2.4 Diagonal Grid Distance Metric

Value of $k = 5$

Method	Hubness	Goodman-Kruskal Index	Accuracy
Original	0.42	0.068	37.08%
SNN	0.34	0.081	31.76%
Local Scaling	-0.17	0.133	38.66%
Mutual Proximity	-0.08	0.105	37.87%

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

[2] Simon Günter and Horst Bunke. Validation indices for graph clustering. *Pattern Recognition Letters*, 24(8):1107–1113, 2003.

[3] Raymond A Jarvis and Edward A Patrick. Clustering using a similarity measure based on shared near neighbors. *Computers, IEEE Transactions on*, 100(11):1025–1034, 1973.

[4] Elias Pampalk, Arthur Flexer, Gerhard Widmer, et al. Improvements of audio-based music similarity and genre classificaton. In *ISMIR*, volume 5, pages 634–637. London, UK, 2005.

[5] 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.

[6] 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.

[7] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *Advances in neural information processing systems*, pages 1601–1608, 2004.