Towards Better Music Recommendation Systems

Group 6: Rishav Raj Agarwal (12577) Saurav Prakash (12642)
Guided by: Dr. Piyush Rai
Indian Institute of Technology. Kanpur

Overview

Motivation

- **Hubs** are data points which keep appearing often as nearest neighbors of large number of other data points.
- Hubness in Music Recommendation is a very active topic of research.

Objective

- Try to create the best recommender using different measures of similarity and scaling.

Dataset [1]

- 10,000 Songs
- 16 features extracted from metadata.

Challenges

- Dataset in form of individual .h5 files for the songs.
- Computational issues handling 10,000 songs.

Methodology

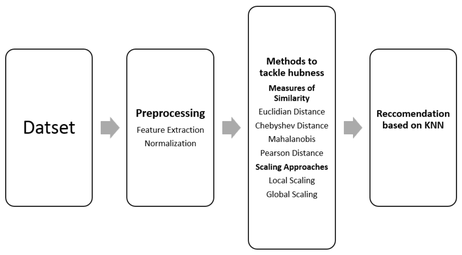

Figure 1: Methodology

Measures of Similarity

Distance	Formula
Euclidian	$\sqrt{\Sigma_k(x_k - y_k)^2}$
Chebyshev	$Max \mid x_k - y_k \mid$
Mahalanobis	$\sqrt{(x-y)\sigma^{-1}(x-y)^T}$
Pearson	$\frac{\Sigma(x,y)}{\sigma_x \sigma_y}$

Table 1: List of Distance functions

Scaling Methods to tackle Hubness

- **Local Scaling**: $LS(d_{x,y}) = exp(-\frac{d^2_{x,y}}{\sigma_x \sigma_y})$
  x and y will be close neighbours only when $d_{x,y}$, is small in comparison to both σ_x and σ_y .
- **Global Scaling**: Transformation of distance matrices to probabilistic mutual proximity (MP) $MP(d_{x,y}) = 1 - P(X < d_{x,y} \cup Y < d_{x,y})$
  The intuition is to increase more closely tie up the objects that have similar nearest neighbourhoods, and repel the objects that have dissimilar neighbourhoods. [2]

Some Definitions

- **Hubness:** Hubness is defined as the skewness of the distribution of k-occurrences N_k
- **Anti Hubs:** objects having a k-occurrence of zero (k = 5).
- **Intrinsic Dimensionality:** The intrinsic dimension is the number of dimensions necessary to represent a data set without loss of information.

Variation with Dimensionality

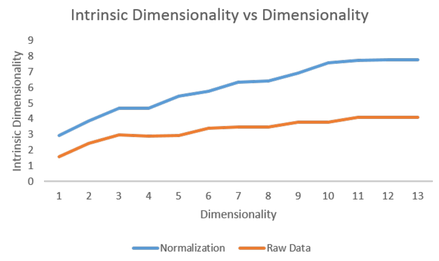

Figure 3: Intrinsic Dimensionality with Dimesionality

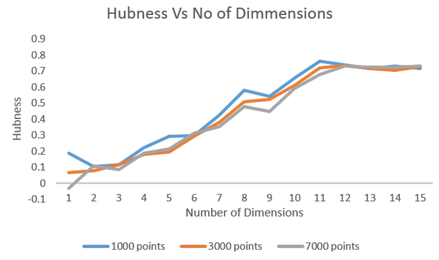

Figure 4: Hubness with Dimesionality

Hubness Analysis

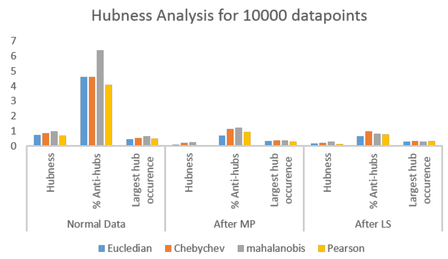

Figure 2: Hubness Analysis for 10000 songs

Results

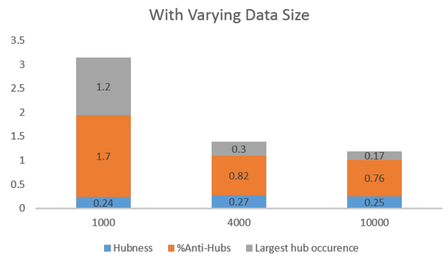

Figure 5: Variation With Datasize after local-scaling

Distance	Decrease Hubness	Improvement
Euclidian	0.57	3.97
Chebyshev	0.63	3.63
Mahalanobis	**0.69**	**5.56**
Pearson	0.58	3.3

Table 2: Improvements on 10000 songs after Local Scaling

Discussion

- **Variation with Dimensionality:** We see that as dimesionality increases, the hubness problem increases. [3]
- **Variation with Data Size:** We see that the hubness and number of anti-hubs decrease with increasing datasize but the changes are very small to be conclusive.
- **Variation With type of Distance Function:** We see that Mahalanobis gives the highest hubness. Points closer to the dataset mean tend to become hubs [2]. As Mahalanobis measures distance number of standard deviations from the mean, we get the best results out of it.
- **Variation with Scaling Method:** We see that Local Scaling Works better for our dataset and gives better recommendations however MP is more effective in reducing Hubness.
- **Other Variations:** We also tried variations of MP like using a Gaussian Distribution to model the probabilities but the results were worse off.

Conclusion

The intrinsic dimensionalty of our dataset is very low (7) , so the hubness problem is not as prominent as we expected. Nevertheless, we were able to significantly improve over the baseline.
The best results we got was using the Mahalanobis Distance with recommendation accuracy was at: 99.18% with a 5.56% improvement due to Local Scaling.

References

[1] Bertin-Mahieux, Thierry, et al. 'The million song dataset.' ISMIR 2011: Proceedings of the 12th International Society for Music Information Retrieval Conference, October 24-28, 2011, Miami, Florida. University of Miami, 2011.

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

[3] RadovanoviĂĆ, MiloÅą, Alexandros Nanopoulos, and Mirjana IvanoviĂĆ. 'Hubs in space: Popular nearest neighbors in high-dimensional data.' The Journal of Machine Learning Research 11 (2010): 2487-2531.