The impact of hubness on music recommendation

Arthur Flexer ARTHUR.FLEXER@OFAI.AT
Austrian Research Institute for Artificial Intelligence, Freyung 6/6, 1010 Vienna, Austria

Abstract

We review the impact of hubness, a general problem of machine learning in high-dimensional spaces, on music recommendation. Due to a problem of measuring distances in high dimensions, hub objects are recommended over and over again while anti-hubs are nonexistent in recommendation lists. After reviewing the theory concerning the hubness phenomenon, we present methods which are able to decisively diminish hubness and its adverse effects in music and general multimedia datasets.

1. Theoretical background

Hubness is a general problem of learning in high-dimensional spaces which has been discovered in music information retrieval (MIR) (Aucouturier & Pachet, 2004), but then gained attention in a general machine learning context where it has been discussed as a new aspect of the curse of dimensionality (Radovanović et al., 2010; Schnitzer et al., 2012). Hub objects appear very close to many other data objects and anti-hubs very far from most other data objects. The effect has been shown to have a negative impact on classification (Radovanović et al., 2010), nearest neighbor based recommendation (Flexer et al., 2012) and retrieval (Schnitzer et al., 2014), outlier detection (Radovanović et al., 2010) and clustering (Tomašev et al., 2014; Schnitzer & Flexer, 2015).

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 (François et al., 2007). Radovanović et al. (2010) presented the argument that for any finite dimensionality, some points are expected to be closer to the center of all data than other points and are at the same time closer, on average, to all other points. Such points closer to the center have a high probability of being hubs, i.e. of appearing in nearest neighbor lists of many other points. Points which are further away from the center have a high probability of being 'anti-hubs', i.e. points that never appear in any nearest neighbor list. It is also important to note that the degree of concentration and hubness is linked to the intrinsic rather than extrinsic dimension of the data space. Whereas 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 (François et al., 2007). Previous research (Schnitzer et al., 2012) has shown that real world data with extrinsic dimensionality as small as 34 can already exhibit the negative effects of hubness.

2. Reducing hubness

In order to reduce hubness and its negative effects, we have proposed two unsupervised methods to re-scale high-dimensional distance spaces (Schnitzer et al., 2012): Local Scaling (LS) and Mutual Proximity (MP). Both methods aim at repairing asymmetric nearest neighbor relations. The asymmetric relations are a direct consequence of the presence of hubs. A hub y is the nearest neighbor of x, but the nearest neighbor of the hub y is another point a ($a \neq x$). This is because hubs are by definition nearest neighbors to very many data points but only one data point can be the nearest neighbor to a hub. The principle of the scaling algorithms is to re-scale distances to enhance symmetry of nearest neighbors. A small distance between two objects should be returned only if their nearest neighbors concur. LS does that by computing a local statistic of the nearest neighbors to rescale the distances. MP assumes that all pairwise distances in a data set follow a certain distribution and computes the mutual probability that two points are true nearest neighbors. Application of LS and MP resulted in a decrease of hubness and an accuracy increase in k-nearest neighbor classification on thirty real world datasets including text, image and music data.

The somewhat related classic shared nearest neighbors

method (SNN) and its impact on hubness have been studied for nearest neighbor classification (Tomašev & Mladenić, 2012) and compared to LS and MP (Flexer & Schnitzer, 2013). It was shown that SNN does reduce hubness but less than LS and MP and that it is not able to improve classification accuracy for all of the data sets used in the study. A different approach to reduce hubness is to center the data either locally or globally (Hara et al., 2015). This gives nearest neighbor classification results comparable to LS and MP when applied to text data. It has also been tried (Schnitzer & Flexer, 2014) to use l^p norms different than the ubiquitous Euclidean l^2 norm for computation of nearest neighbor lists to avoid the negative effects of hubness. Comparison of using different l^p norms to LS, MP and SNN shows (Flexer & Schnitzer, 2015) that it is highly problem dependent which of the approaches works best.

3. Impact on music recommendation

We now review results (Schnitzer et al., 2012) which we have achieved by applying MP to improve the quality of a real world music discovery system: the FM4 Soundpark[1]. The FM4 Soundpark is a web platform run by the Austrian public radio station FM4 where artists can upload and present their music free of charge. Visitors of the website can listen to and download all the music at no cost, with most recent uploads being displayed at the top of the web-site. To allow a more intuitive and appealing access to the full database regardless of publication date of a song, we implemented a recommendation system using a content-based music similarity measure (Gasser & Flexer, 2009). The visualization displays an incrementally constructed nearest neighbor graph showing the five most similar songs to the currently playing one (see Fig. 1). High hubness causes some songs to never occur in the nearest neighbor lists at all, since hubs crowd the nearest neighbors lists and are being recommended repeatedly. As a result only 72.6% of the songs are reachable in the recommendation interface, i.e. over a quarter of songs are never recommended. Further analysis of the nearest neighbor graph shows that only less than a third of the songs are likely to be recommended, since only those are part of one large strongly connected subgraph. Application of MP is able to increase reachability to 86.2% while at the same time improving retrieval accuracy. Random sampling of five songs from the ten nearest neighbors computed via MP increases this to 93.7% while still showing good retrieval accuracy.

In a large meta-study analyzing 17 algorithms from an evaluation campaign on "Audio Music Similarity and Retrieval" (Flexer et al., 2012), we were able to show that many different approaches to compute audio similarity based recommendations are negatively impacted by hubness. We were also able to show that hub songs, when being recommended as being very similar, are judged to be less perceptually meaningful than non-hub songs by human evaluators. Again these negative effects could be reduced by applying MP to re-scale the distances. We also investigated the impact of hubness on general multimedia retrieval by analyzing textual and image data (Schnitzer et al., 2014). Negative effects were similar to those observed for music data, with the amount of unreachable objects ranging from about 20% to over 50%. In the case of a data set comprised of twitter messages, one single message appeared in the nearest neighbor lists (size five) of more than a quarter of all messages. Again both LS and MP are able to decisively improve this situation. And finally, multimedia recommendation via collaborative filtering has also been shown to suffer from the hubness problem which can also be mitigated by applying MP (Knees et al., 2014).

Figure 1. Music recommendation based on visualization of incrementally constructed five-nearest neighbor graph.

4. Conclusion

The main impact of hubness on music recommendation and discovery, but also on general multimedia retrieval, is the dominance of hub objects in nearest neighbor based recommendation lists which at the same time causes anti-hub objects never to be retrieved. The review of work on reducing these negative effects suggests to make hubness analysis an integral part when building a recommendation system.

Acknowledgments

This research is supported by the Austrian Science Fund (FWF, project P27082).

[1] http://fm4.orf.at/soundpark

References

Aucouturier, Jean-Julien and Pachet, François. Improving timbre similarity: How high is the sky? *Journal of Negative Results in Speech and Audio Sciences*, 1(1), 2004.

Flexer, Arthur and Schnitzer, Dominik. Can shared nearest neighbors reduce hubness in high-dimensional spaces? In *Proceedings of 1st International Workshop on High Dimensional Data Mining (HDM), in conjunction with the IEEE International Conference on Data Mining (IEEE ICDM 2013)*, pp. 460–467, Dec 2013.

Flexer, Arthur and Schnitzer, Dominik. Choosing l^p norms in highdimensional spaces based on hub analysis. *Neurocomputing*, 2015. doi: http://dx.doi.org/10.1016/j.neucom.2014.11.084. to appear, available online 18 April 2015.

Flexer, Arthur, Schnitzer, Dominik, and Schlüter, Jan. A mirex meta-analysis of hubness in audio music similarity. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 175–180, 2012.

François, Damien, Wertz, Vincent, and Verleysen, Michel. The concentration of fractional distances. *IEEE Trans. on Knowledge and Data Engineering*, 19:873–886, 2007.

Gasser, Martin and Flexer, Arthur. FM4 SoundPark: audio-based music recommendation in everyday use. In *Proceedings of the 6th Sound and Music Computing Conference*, pp. 23–25, 2009.

Hara, Kazuo, Suzuki, Ikumi, Shimbo, Masashi, Kobayashi, Kei, Fukumizu, Kenji, and Radovanović, Miloš. Localized centering: Reducing hubness in large-sample data. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pp. 2645–2651. AAAI Press, 2015.

Knees, Peter, Schnitzer, Dominik, and Flexer, Arthur. Improving neighborhood-based collaborative filtering by reducing hubness. In *Proceedings of International Conference on Multimedia Retrieval*, ICMR 2014, pp. 161–168, 2014.

Radovanović, Miloš, Nanopoulos, Alexandros, and Ivanović, Mirjana. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531, 2010.

Schnitzer, Dominik and Flexer, Arthur. Choosing the metric in high-dimensional spaces based on hub analysis. In *22th European Symposium on Artificial Neural Networks, ESANN 2014, Bruges, Belgium, April 23-25*, 2014.

Schnitzer, Dominik and Flexer, Arthur. The Unbalancing Effect of Hubs on K–medoids Clustering in High-Dimensional Spaces. In *Proceedings of the International Joint Conference on Neural Networks*, 2015.

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

Schnitzer, Dominik, Flexer, Arthur, and Tomašev, Nenad. A case for hubness removal in high-dimensional multimedia retrieval. In *Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014*, pp. 687–692, 2014.

Tomašev, Nenad and Mladenić, Dunja. Hubness-aware shared neighbor distances for high-dimensional k-nearest neighbor classification. In *Hybrid Artificial Intelligent Systems*, pp. 116–127. Springer, 2012.

Tomašev, Nenad, Radovanović, Miloš, Mladenić, Dunja, and Ivanović, Mirjana. The role of hubness in clustering high-dimensional data. *IEEE Transactions on Knowledge and Data Engineering*, 26(3):739–751, 2014.