Hubness in the Context of Feature Selection and Generation (Extended Abstract)

Miloš Radovanović
Department of Mathematics and Informatics
University of Novi Sad
Serbia
radacha@dmi.uns.ac.rs

Alexandros Nanopoulos
Institute of Computer Science
University of Hildesheim
Germany
nanopoulos@ismll.de

Mirjana Ivanović
Department of Mathematics and Informatics
University of Novi Sad
Serbia
mira@dmi.uns.ac.rs

Hubness is a property of vector-space data expressed by the tendency of some points (hubs) to be included in unexpectedly many k-nearest neighbor (k-NN) lists of other points in a data set, according to commonly used similarity/distance measures. Alternatively, hubness can be viewed as increased skewness of the distribution of node in-degrees in the k-NN digraph obtained from a data set. Hubness has been observed in several communities (e.g., audio retrieval), where it was described as a problematic situation. The true causes of hubness have, for the most part, eluded researchers, which is somewhat surprising given that the phenomenon represents a fundamental characteristic of vector-space data.

In recent work, we have shown that hubness is actually an inherent property of data distributions in multidimensional space, caused by high intrinsic dimensionality of data. Hubness can therefore be viewed as a notable novel aspect of the "curse of dimensionality." We have explored the implications of hubness on various tasks, including distance-based methods for machine learning [1], and vector space models for information retrieval [2], with the common conclusion that hubness is an important concern when dealing with data that is intrinsically high-dimensional.

We have concentrated our research efforts so far on explaining the origins of hubness and its effects on different tasks, assuming a given set of features defined for a particular data set. Although our works briefly consider the interaction of hubness with dimensionality reduction, the implications of hubness on feature selection, and especially generation, are still open research questions.

Besides hubness, we will center our current discussion on the notion of the *cluster assumption* (CA) which, assuming that data contains labels, roughly states that two points in the same cluster should in most cases be of the same class. This assumption is one of the pillars of semi-supervised ML methods, and is also known in IR circles as the *cluster hypothesis*, which is formulated in an analogous manner using the notion of relevance. The cluster assumption, i.e., the degree of its violation, effectively represents the degree to which the featural representation of the data fails to correspond with some notion of "ground truth" about the data given, e.g., by class labels. A high degree of CA violation indicates that models will be difficult to build from the data, and may suggest a reconsideration of the featural representation.

For a given data set and distance measure, let $N_k(\mathbf{x})$ denote the number of times point $\mathbf{x}$ occurs in the k-NN lists of other points in the data set. We express hubness using the skewness of the distribution of $N_k(\mathbf{x})$, as its standardized third moment, denoted S_{N_k}. Also, let $BN_k(\mathbf{x})$ be the number of times $\mathbf{x}$ occurs in the k-NN lists of other points, where the labels of $\mathbf{x}$ and the points in question do not match (making $BN_k(\mathbf{x})$ a measure of "badness" of $\mathbf{x}$). The normalized sum of $BN_k(\mathbf{x})$ for a given data set, BN_k ratio, represents one way to express the degree of violation of the CA.

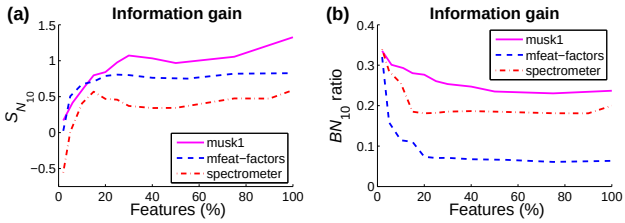

Figure 1: Skewness and "badness" ratio.

Figure 1 illustrates how S_{N_k} and BN_k ratio change when features are selected using the classical information gain method, on three data sets from the UCI repository also used in [1]. Regarding Fig. 1(a), looking from right to left, skewness of N_k stays relatively constant until a small percentage of the original number of features is left, when it abruptly drops. This is the point where the intrinsic dimensionality is reached, with further selection incurring loss of information. This loss is also visible in Fig. 1(b), where at similar points there is an increase in BN_k ratio, suggesting that the reduced representation ceases to reflect the information provided by labels very well.

Observing the two charts in the opposite direction, from left to right, offers a glimpse into the benefits and drawbacks of feature *generation*. Adding features that bring new information to the data representation will ultimately increase S_{N_k} and produce hubs. Furthermore, for the chosen examples, the reduction of BN_k ratio "flattens out" fairly quickly, limiting the usefulness of adding new features in the sense of being able to express the "ground truth." Depending on the application, instead of BN_k ratio some other criterion could have been used in Fig. 1(b), like classifier error rate, producing similarly shaped curves. While the majority of research in feature selection/generation has focused on optimizing criteria reminiscent to those in Fig. 1(b), little attention has been paid to the fact that in intrinsically high-dimensional data hubness will result in an uneven distribution of the cluster assumption violation (in our case, hubs will generally attract more label mismatches with neighboring points), and with it an uneven distribution of responsibility for classification or retrieval error among data points. We believe that investigating the interaction between hubness and different notions and analogues of CA violation can result in important new insights relevant to the tasks of feature selection and generation.

REFERENCES

[1] M. Radovanović, A. Nanopoulos, and M. Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*. Forthcoming.

[2] M. Radovanović, A. Nanopoulos, and M. Ivanović. On the existence of obstinate results in vector space models. In *Proc. SIGIR*, 2010.