ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Choosing the Metric in High–Dimensional Spaces Based on Hub Analysis

Dominik Schnitzer and Arthur Flexer *

Austrian Research Institute for Artificial Intelligence (OFAI)
Freyung 6/6, 1010 Wien – Austria

Abstract. To avoid the undesired effects of distance concentration in high-dimensional spaces, previous work has already advocated the use of fractional ℓ^p norms instead of the ubiquitous Euclidean norm. Closely related to concentration is the emergence of hub and anti-hub objects. Hub objects have a small distance to an exceptionally large number of data points while anti-hubs lie far from all other data points. The contribution of this work is an empirical examination of concentration and hubness, resulting in an unsupervised approach for choosing an ℓ^p norm by minimizing hubs while simultaneously maximizing nearest neighbor classification.

1 Introduction and Related Work

This work examines fractional ℓ^p norms in high-dimensional spaces in the context of the problem of hubs. A number of publications [1, 2, 3] have recently focused on the emergence of hubs as a new aspect of the curse of dimensionality [4]. Hubs have an exceptionally low distance to a high number of objects and therefore are nearest neighbors of an exceptionally large percentage of other data points. As a result, other objects (anti-hubs) are pushed out of all nearest neighbor lists. It was shown that this behavior has a negative impact on many machine learning tasks including classification [1], nearest neighbor based recommendation [5, 6], outlier detection [1, 7] and clustering [8].

The hubness problem is closely linked to the concentration of distances in high-dimensional data spaces [1]. Concentration is the surprising characteristic of all points in a high-dimensional space to be at almost the same distance to all other points in that space [9]. It is usually measured as a ratio between spread and magnitude, e.g. the ratio between the standard deviation of all distances to an arbitrary reference point and the mean of these distances. If the standard deviation stays more or less constant with growing dimensionality while the mean keeps growing, the ratio converges to zero with dimensionality going to infinity. In such a case the distance contrast decreases and it is said that the distances concentrate. Proofs concerning concentration of distances and all points being at the same distance to all other points have been formulated for dimensionality approaching infinity. Radovanović et al. [1] presented the argument that in the finite case, some points are expected to be closer to the center 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

*This research is supported by the Austrian Science Fund (FWF): P24095.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

from the center have a high probability of being 'anti-hubs', i.e. points that never appear in any nearest neighbor list. This was evaluated for cosine and Euclidean (ℓ^2) norm on real world data but also observed for $\ell^{0.5}$ using i.i.d. normal and uniform data. 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.

The concentration effect was studied by Aggarwal et al. [10] for Euclidean and fractional ℓ^p norms. In fact Aggarwal et al. come to the conclusion that from a theoretical and empirical perspective the Euclidean (ℓ^2) norm is often not the preferred metric for high-dimensional data mining applications since fractional norms provide a higher distance contrast. Experiments also show that choosing the right fractional norm as opposed to the Euclidean norm could significantly improve the effectiveness of standard k–nearest neighbor (kNN) classification in high-dimensional spaces. This observation was more closely investigated by François et al. [11] who follow a supervised approach to infer the optimum ℓ^p norm using labeled training data.

Our work pursues the ideas of Aggarwal et al. [10] and François et al. [11] in the light of the effects of hubs and anti-hubs. We show empirically that the degree of hubs and anti-hubs in a data set can help selecting the optimum ℓ^p norm. Based on these results we propose a fully unsupervised approach for choosing an ℓ^p norm which maximizes nearest neighbor classification.

2 Methods and Data

Hubs and anti-hubs are found by looking at all kNN lists of a data set X. For a given neighborhood size k, the k-occurrence ($O^k(x)$) of a point $x \in X$ is then computed by counting the number of occurrences of x in the kNN of each point $x_i \in X, x_i \neq x$. Using O^k we then define hubs (H^k) and anti-hubs (A^k) as:

$$A^k = \left\{a \in X | O^k(x) = 0\right\}, \qquad H^k = \left\{h \in X | O^k(x) \geq 2k\right\}.$$

Anti-hubs (A) never occur in the kNN, i.e. have a O^k of zero, while hubs (H) occur equal or more than twice as often ($2k$) as expected. To asses the overall impact of hubness in a data set Radovanović et al. [1] proposed to compute 'hubness' (S^k) which he defined as the skewness of the histogram of the O^k. The higher the measured sample skewness of the O^k histogram, the higher the impact of hubs in the kNN:

$$S^k = \frac{\mathrm{E}\left[(O^k - \mu_{O^k})^3\right]}{\sigma_{O^k}^3}.$$

We use this measure to identify high-dimensional data sets showing strong hubness in the Euclidean space by choosing data sets where $S^{k=5} > 2$.[1] The data sets identified are: *Protein*, *Splice*, *Gisette* and *Dexter* from the UCI machine

[1]Methods for hubness data analysis are available in our Matlab hub–toolbox: `http://www.ofai.at/research/impml/projects/hubology.html`

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

learning archive [12], two standard image-classification data sets (*Leeds Butterfly* [13], *17 Flowers* [14]) and a data set from the text-retrieval domain, *Twitter (C1ka)* [15]. The dimensionality *dim*, size of data set n and hubness $S^{k=5}$ of the original Euclidean space is listed in Table 1. Data sets are used as they are available on their respective websites without any additional normalization. The extrinsic dimensionality ranges from 60 (*Splice*) to 49 820 (*Twitter (C1ka)*) while the measured hubness goes from rather moderate values of 2.9 (*Gisette* and *Dexter*) to extreme values of 43.1 (*Protein*) in ℓ^2.

Like Aggarwal et al. [10] we will evaluate the impact of changing the ℓ^p norm by reporting the kNN classification accuracy using leave-one-out cross-validation. The classification is performed via a majority vote among the k nearest neighbors, with the class of the nearest neighbor used for breaking ties. We denote the kNN accuracy as C^k. In the context of a retrieval problem, higher values would indicate better retrieval quality.

3 Experiments and Results

To measure the impact of hubs and anti-hubs on a given data set we propose two measures (i) anti-hub occurrence (A^k_{occ}) and (ii) hub occurrence (H^k_{occ}). Whereas A^k_{occ} is the percentage of data points that act as anti-hubs, H^k_{occ} is the percentage of hub points in all kNN lists. We include these measures in our experiments to evaluate a given ℓ^p norm in terms of anti-hubs and hubs at a selected neighborhood radius k:

$$A^k_{occ} = \frac{1}{|X|}\left|A^k\right|, \qquad H^k_{occ} = \frac{1}{|X|}\sum_{h\in H^k}\frac{O^k(h)}{k}.$$

We do not use the hubness measure (S^k), i.e. the skewness of the O^k, since it does not equally account for hubs and anti-hubs in the measurements. By computing the sample skewness, hubs with a theoretical maximum $O^k(h) = |X| - 1$, have a much higher influence on the measure than anti-hubs since their difference to the μ_{O^k} contributes to S^k to the third power. Additionally our experiments with S^k in this context did not show a smooth but oscillating change of values when stepping through the ℓ^p norms, making S^k unfit for our purpose.

To investigate the relation of hubs and anti-hubs to a certain ℓ^p norm we compute A^k_{occ} and H^k_{occ} for our selected data sets. We set our neighborhood size to $k = 1$ (i.e., we only look at each point's nearest neighbor) while changing the ℓ^p norm from $p = 0.25, 0.5, 0.75, \ldots, 4$. For each step in p we compute the kNN classification rate $C^{k=5}$. Figure 1 plots the results for each of the selected data sets. A^k_{occ} is plotted in the first column of the figures, H^k_{occ} in the second column and the classification rate $C^{k=5}$ in the third column of the figures. Each of the measures is computed while varying the p as discussed. Please note that results using a larger neighborhood size to compute A^k_{occ} and H^k_{occ} or with one nearest neighbor classifciation ($C^{k=1}$) did not substantially change the following results.

Looking at the figures we first note a very high similarity between the anti-hub (A^k_{occ}) and hub (H^k_{occ}) curves. This behavior is expected as a higher number

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

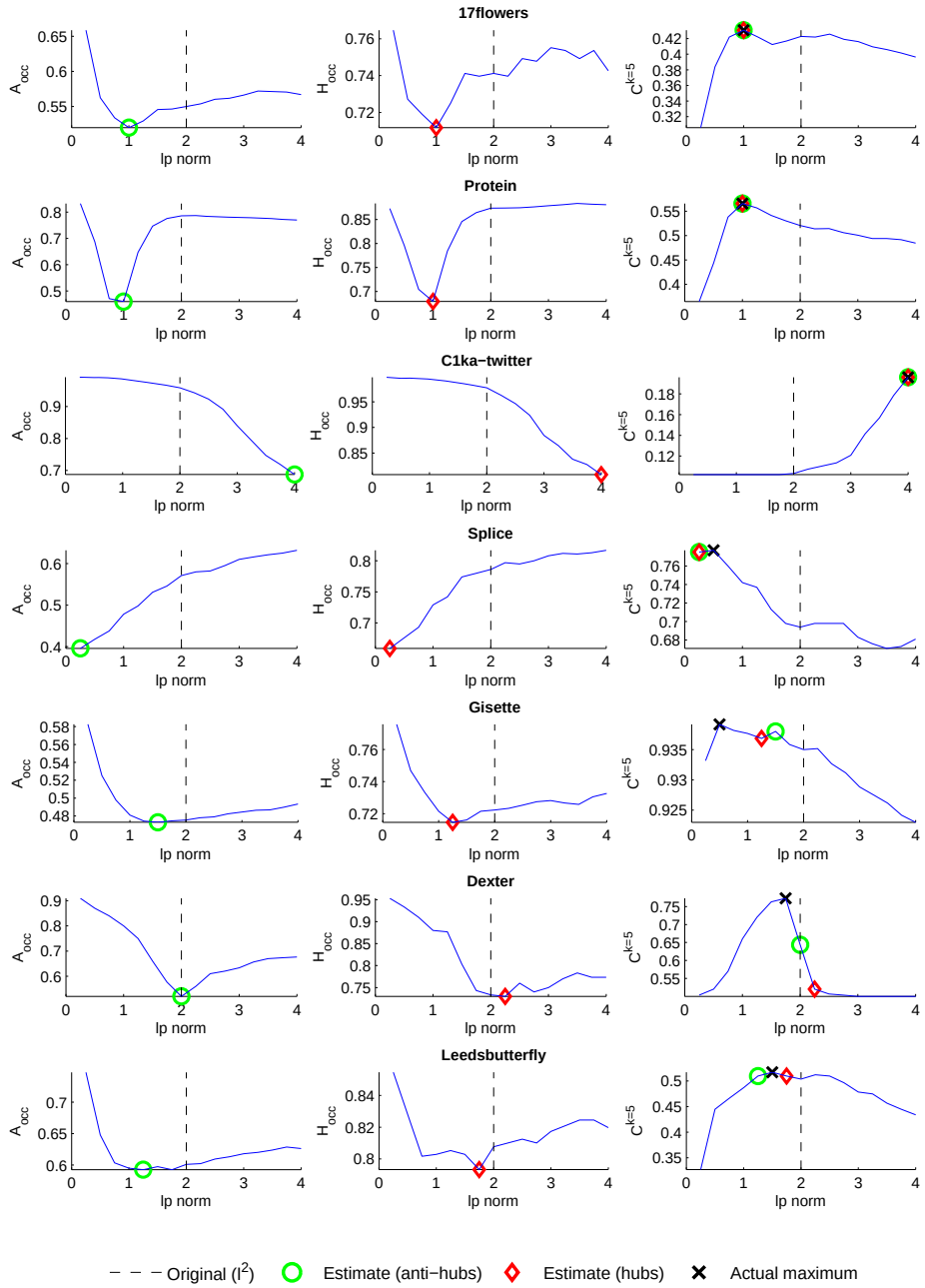

Fig. 1: The minimum in anti-hub (A^k_{occ}) and hub (H^k_{occ}) occurrence while changing the ℓ^p norm is closely related to the maximum kNN classification rate (C^k).

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

dataset	*dim*	n	$S^{k=5}$	*original*		*max.* C^k		*estimated max. using* A^k_{occ}		*estimated max. using* H^k_{occ}	
				ℓ^p	$C^{k=5}$	ℓ^p	$C^{k=5}$	ℓ^p	$C^{k=5}$	ℓ^p	$C^{k=5}$
Dexter	20 000	300	2.9	2	64.3%	1.75	**77.3%**	2	**64.3%**	2.25	52.0%
Gisette	5 000	6 000	2.9	2	93.5%	0.5	**93.9%**	1.5	**93.8%**	1.25	**93.7%**
Leeds Butt.	36 000	832	3.5	2	50.4%	1.5	**51.7%**	1.25	**51.0%**	1.75	**51.0%**
17 Flowers	36 000	1 360	3.9	2	42.3%	1	**43.1%**	1	***43.1%**	1	***43.1%**
Splice	60	1 000	5.6	2	69.4%	0.5	**77.7%**	0.25	**77.5%**	0.25	**77.5%**
Twitter	49 820	969	14.6	2	10.3%	4	**19.6%**	4	***19.6%**	4	***19.6%**
Protein	357	6 621	43.1	2	52.1%	1	**56.6%**	1	***56.6%**	1	***56.6%**

Table 1: Data sets, their dimensionality *dim* and size n, classification rates (C^k) in the original Euclidean space (ℓ^2), possible maximum (*max.* C^k) and estimated maximum ℓ^p using A^k_{occ} and H^k_{occ}. Better or equal C^k when compared to the original data in bold, an asterisk indicates that respective methods were able to find the maximum.

of objects not occurring in the kNN at all has to lead to higher O^k values for the remaining objects. Additionally the change in the kNN classification accuracy (C^k) seems to be in accordance with Aggarwal et al. [10] with the values peaking at $p \neq 2$. Furthermore and more interestingly the peak in C^k seems to concur with either A^k_{occ} or H^k_{occ} being at or close to their minimum. In view of the fact that neither the computation of A^k_{occ} nor H^k_{occ} include any knowledge about classes these empirical results give a strong argument that both measures could be effective for choosing the optimum ℓ^p norm.

Table 1 summarizes the results. In the table we list the original kNN classification rate (C^k) in ℓ^2, the actual maximum and the two estimated maxima using A^k_{occ} and H^k_{occ}. In three data sets (*17 Flowers*, *Protein* and *Twitter (C1ka)*) we are able to identify the best ℓ^p norm according to C^k by using the minima of both A^k_{occ} or H^k_{occ}. The increase in C^k ranges from 0.9 to 9.3 percentage points. The optimum norm is twice ℓ^1 and once ℓ^4. In three further cases (*Splice*, *Gisette* and *Leeds Butterfly*) both measures are able to identify a better ℓ^p norm than the Euclidean base case, but closely fail to identify the possible maximum. The increase in C^k ranges from 0.4 to 8.1 percentage points. In the case of *Dexter* and by using H^k_{occ} ($p = 2.25$) as decision, the proposed method would lead to a drop in classification accuracy by 12.3 percentage points. Using A^k_{occ} however would stay with the Euclidean norm, thus suggesting no change of norm. The theoretical maximum is at $p = 1.75$. Upon closer inspection of the results, we see H^k_{occ} closely missed ℓ^2 because a single hub occurrence ($O^k(h)$) is increased by a count of 1 (and the theoretical C^k maximum is missed due to an increase of 4 counts). The small data set size ($|X| = 300$) could be the cause for this result.

4 Summary

This work linked finding the optimum ℓ^p norm (in terms of kNN classification rates) for high-dimensional data to hubs and anti-hubs occurring in high-

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

dimensional data. In an empirical study we presented strong evidence that the optimum ℓ^p norm for data sets with high hubness in the Euclidean space can be found at values of p, where hubs and anti-hubs have their minimal impact on the data. To identify these points we propose to measure the hub (H^k_{occ}) or anti-hub (A^k_{occ}) occurrence as defined in this work. Using these measures we were able to identify better norms in six of the seven analyzed data sets.

References

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

[2] Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, Yuji Matsumoto, and Marco Saerens. Investigating the effectiveness of laplacian-based kernels in hub reduction. In *Proceedings of the Twenty-Sixth Conference on Artificial Intelligence (AAAI)*, pages 1112–1118, 2012.

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

[4] Richard Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.

[5] Dominik Schnitzer, Arthur Flexer, Markus Schedl, and Gerhard Widmer. Using mutual proximity to improve content-based audio similarity. In *ISMIR*, pages 79–84, 2011.

[6] Alexandros Nanopoulos, Miloš Radovanović, and Mirjana Ivanović. How does high dimensionality affect collaborative filtering? In *Proceedings of the third ACM conference on Recommender systems*, pages 293–296. ACM, 2009.

[7] Arthur Flexer and Dominik Schnitzer. Using mutual proximity for novelty detection in audio music similarity. In *6th International Workshop on Machine Learning and Music (MML), In conjunction with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*, Prague, Czech Republic, 2013.

[8] Nenad Tomašev, Miloš Radovanović, Dunja Mladenić, and Mirjana Ivanović. The role of hubness in clustering high-dimensional data. In *Advances in Knowledge Discovery and Data Mining*, pages 183–195. Springer, 2011.

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

[10] Charu Aggarwal, Alexander Hinneburg, and Daniel Keim. On the surprising behavior of distance metrics in high dimensional space. In *Database Theory - ICDT 2001*, Lecture Notes in Computer Science, pages 420–434. Springer Berlin/Heidelberg, 2001.

[11] Damien François, Vincent Wertz, and Michel Verleysen. Choosing the metric: A simple model approach. In *Meta-Learning in Computational Intelligence*, volume 358 of *Studies in Computational Intelligence*, pages 97–115. Springer Berlin Heidelberg, 2011.

[12] K. Bache and M. Lichman. UCI machine learning repository, 2013.

[13] Josiah Wang, Katja Markert, and Mark Everingham. Learning models for object recognition from natural language descriptions. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2009.

[14] M-E Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Computer Vision, Graphics & Image Processing, 2008. ICVGIP'08. Sixth Indian Conference on*, pages 722–729. IEEE, 2008.

[15] Markus Schedl. On the Use of Microblogging Posts for Similarity Estimation and Artist Labeling. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, the Netherlands, August 2010.