EXPLORING THE HUBNESS-RELATED PROPERTIES OF OCEANOGRAPHIC SENSOR DATA

Nenad Tomašev, Dunja Mladenić
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
e-mail: nenad.tomasev@ijs.si, dunja.mladenic@ijs.si

ABSTRACT

In this paper we examine how the high dimensionality of oceanographic sensor data impacts the potential use of nearest-neighbor machine learning methods. We focus on one particular consequence of the curse of dimensionality – hubness. We examine the hubness of oceanographic data and show how it can be used to visualize and detect both prototypical sensors/locations, as well as ambiguous and potentially erroneous ones. We proceed to define an easy classification problem on the data, showing that the recently developed hubness-aware classification methods may help to overcome some of the hubness-related issues in sensor data.

1 INTRODUCTION

Various sensor arrays spread across the world have endowed us with a greater insight into the dynamics of many natural phenomena. Due to the sheer quantity of such data, semi-automatic analysis and interpretation is not an option, but a necessity. Machine learning methods are hence used for prediction, categorization, clustering, error-detection and other tasks which may prove beneficial within a certain problem context.

Nearest neighbor methods are frequently used in machine learning. They are based on an intuitive notion that similar instances (where similarity is measured by some appropriate metric) often share some common properties. Therefore, in order to deduce something about the current point of interest, its nearest neighbors can be examined and used to infer the desired property. If this property is in fact the instance label, we could use one of the many *k*-nearest neighbor classification methods proposed in literature.

The basic *k*-nearest neighbor method (*k*NN) was first introduced in [1]. It is a simple procedure, where a majority vote over the *k*-nearest neighbor set is taken in order to determine the label of the point of interest. This simple rule was shown to have some useful asymptotic properties [2][3] and hence became very popular – many extensions of the basic algorithm have been successfully applied to various problems. Some very robust general-purpose *k*-nearest neighbor methods have been proposed recently, such as [4] where the metric is learned from the data in such a way that in the new imposed metric space – the proportion of neighbors in *k*-neighbor sets which are of the same class as the observed points is maximized.

Most real world data these days is inherently high-dimensional, whether its images, text, medical data, or – time series data, such as sensor data addressed in this paper. It was shown that high-dimensional data are likely to express significant *hubness* [5]. In such cases, some very influential points emerge (*hubs*) which greatly impact all aspects of nearest-neighbor reasoning. Hubness will be discussed in more detail throughout Section 2.

Ever since the topic of climate change started gaining focus, analyzing such sensor data is becoming more important. Therefore, we wished to provide further insights into the applicability of nearest-neighbor reasoning in such data. Section 3 provides description of the data we have used in the experiments.

We examined the *k*-occurrence distributions based on several measured physical quantities. We used these results to map influential sensor nodes and to detect those which might pose severe difficulties in subsequent nearest-neighbor based inference. We also experimented with several recently developed hubness-aware classification methods and tested their applicability to the problem domain. Experimental results are presented in Section 4.

2 HUBNESS

The *curse of dimensionality* is a term commonly used to address the difficulties inherent in dealing with such data in various practical applications. One of these difficulties is known as *hubness*. Under hubness, different points occur in *k*-neighbor sets with increasingly unequal frequencies. Some points occur in many *k*NN sets, while others occur either very rarely or not at all. The former are referred to as *hubs* and the latter as *anti-hubs.* More specifically, hubness refers to an increasing *skewness* (the third central moment, which describes asymmetry) in the *k*-occurrence distribution in high-dimensional data [5]. This property of the *k*-occurrence distribution was successfully used in [10] to define a hubness-based clustering algorithm aimed specifically at clustering high-dimensional data. This shows that even such detrimental phenomena can be used to our advantage if understood properly.

2.1 BAD HUBNESS

When the data is labeled (i.e. meaningful categories exist), it is possible to distinguish between two sorts of *k*-occurrences: the *good* and the *bad* occurrences (implying *good hubs* and *bad hubs*, respectively). The distinction is made based on the nature of their influence on *k*NN classification. When a neighbor shares the same label as the observed point of interest, that is a *good occurrence*. Label mismatches define *bad occurrences* and add to the *bad hubness* of the neighbor point. Hence, the total number of *k*-occurrences of point x ($N_k(x)$) can be decomposed into the sum of its good and bad occurrences: $N_k(x) = GN_k(x) + BN_k(x)$. Bad hubness can be expected in border regions between different categories, as well noisy, erroneous, or otherwise mislabeled data. Some bad hubness, however, is no more than a consequence of high-data dimensionality and class imbalance.

Hubness-aware classification methods aim at diminishing the influence of bad hubness or otherwise exploiting it in other ways.

2.2 HUBNESS-AWARE CLASSIFICATION

There are several ways in which hubness in general and bad hubness specifically can be dealt with. The simplest approach was suggested in [5], where hubness-based weights were incorporated into the *k*NN rule in order to reduce the weight of votes from bad hubs, since these points were considered essentially unreliable. We will refer to this algorithm as hw-*k*NN.

This was taken a step further in [6], where *class-hubness* was considered instead. The algorithm itself was based on the fuzzy-nearest neighbor framework [7]. So, instead of decomposing the total of *k*-occurrences into good and bad hubness, it was deemed more beneficial to simply take all the occurrence information into account by treating these occurrence probabilities as fuzzy neighbor votes. A Bayesian alternative to the fuzzy approach was introduced in [8], where a simple, easily extensible Naïve Bayesian framework for probabilistic *k*NN classification was presented. An information-theoretic approach was the most recent among the algorithms relying on class-hubness for hubness-aware *k*NN classification [9]. We will refer to these algorithms as h-FNN, NHBNN and HIKNN, respectively.

2.3 HUBNESS IN TIME-SERIES DATA

The hubness phenomenon in time series data in general has recently received some attention [11]. It was shown that, even though time series data do not usually exhibit excessively high dimensionality, it often leads to some tangible hubness. The hubness-weighted *k*NN algorithm has been thoroughly tested on this data and shown to lead to more accurate classification when combined with the DTW (dynamic time warping) distance [12]. It will, however, become apparent later in this paper that hw-*k*NN may not be the best hubness-aware approach for nearest-neighbor time series classification, at least in the oceanographic domain.

3 THE OCEANOGRAPHIC SENSOR DATA

In our experiments, we were working with the Integrated Ocean Observing System data (http://www.ioos.gov/). We were analyzing a sample of measurements from many nodes and attached sensors in a period of 20 days in November 2010. Each sensor is monitoring some physical quantity. We analyzed 8 such quantities: air temperature, barometric pressure, wind observation, water level observation, water level prediction, salinity, water temperature and conductivity. The data came from sensors distributed across the coastlines of North America, so it was partly about the Pacific, partly about the Atlantic ocean and also partly about the Great Lakes. These three location profiles we used as the labels for the sensors, thereby dividing them into 3 location-categories. Each physical property was analyzed separately.

There were some missing values in the data, but not much. Out of the total 4801 time points, usually 50-100 was missing, sometimes none. The values were sampled once every six minutes. This means that there was essentially little difference between neighboring points, so we replaced the missing values by the means of the closest known values.

4 EXPERIMENTS

In our experiments we used two distance measures: the Manhattan distance (sum of absolute differences) and the variance of the difference between the two series over time. The basic hubness-related properties for the data are given in Table 1, separately for the two distance measures.

Two of the sensor types were only present in one region (so they were all of the same label) – conductivity and salinity and have hence not been included in subsequent classification tests. We see that both the *k*-occurrence distribution skewness and the bad hubness are quite similar in both metrics. Most sensor-type data sets exhibit medium skewness, which is consistent with observations from [11], while two measurement types also exhibited quite high hubness, as well as bad hubness in particular: wind and water temperature measurements. The two metrics producing similar results, we will only show the experiments on the Mahnhattan metric in the classifier tests.

Sensor type	size	S_{N3}	BN_3	S_{N5}	BN_5
air temperature	211	0.34	4.7%	0.14	6.7%
barometric pressure	214	0.26	3.4%	-0.06	4.2%
wind	205	3.8	23%	3.6	28%
water level obs.	238	0.6	8.1%	0.47	10%
water level pred.	218	0.34	8.7%	-0.03	11%
salinity	18	-0.13	-	-0.67	-
water temperature	183	0.81	22%	0.67	26%
conductivity	18	0	-	-0.73	-
air temperature	211	0.60	6%	0.55	7.9%
barometric pressure	214	0.11	3.9%	-0.05	4.3%
wind	205	5.2	20%	4.8	24%
water level obs.	238	0.92	9.5%	0.92	12%
water level pred.	218	0.27	6.6%	-0.03	8.9%
salinity	18	0.79	-	0.68	-
water temperature	183	1.16	26%	1.40	31%
conductivity	18	1.01	-	0.81	-

Table 1: *The summary of the data: the number of sensors of a given type, skewness of the 3-NN and 5-NN occurrence distribution (S_{N3}, S_{N5}) and bad hubness of the respective distributions (BN_3, BN_5). The upper part of the table represents results for the Manhattan metric, the lower part for the between-series difference variance.*

It is possible to visualize these bad hubs which are expected to exhibit detrimental influence on nearest-neighbor reasoning in this data. This is shown in Figures 1 and 2, where each sensor was mapped onto a world map according to its latitude/longitude. The size of the circles is proportional to sensor hubness (so, big circles correspond to prototypical, influential points) – and the shade/color to bad hubness (the darker circles corresponding to bad hubs). Figures 1 and 2 represent the two sensor types which were found to exhibit hubness and bad hubness in particular. Two things are apparent from these visualizations. First of all, there are some very big bad hubs. Not only do these sensors exhibit bad hubness, they exhibit it quite frequently. Also, in Figure 1 we see that these big bad hubs are located amidst some low-hubness *good* points, which exhibit no bad hubness. Since both the wind and the water temperature profiles are expected to be similar between these good and bad points, we can conclude that the reasons for bad hubness might even be artificial in this case, like some measurement equipment malfunctioning and producing noisy data. This suggests that bad hubness might also be used for potential error-detection (even though, obviously – not all bad hubs are erroneous data points – and not all erroneous points exhibit bad hubness).

We have tested all the existing hubness-aware classification methods described in Section 2.2 and have compared them to *k*NN as the baseline. The Manhattan metric was used in the experiments. The classification accuracies were obtained via 10-times-10-fold cross validation procedure. Corrected resampled *t*-test was used to test statistical significance. The results for two fixed *k*-values are shown in Table 2. Default parameter settings were used for each of the algorithms.

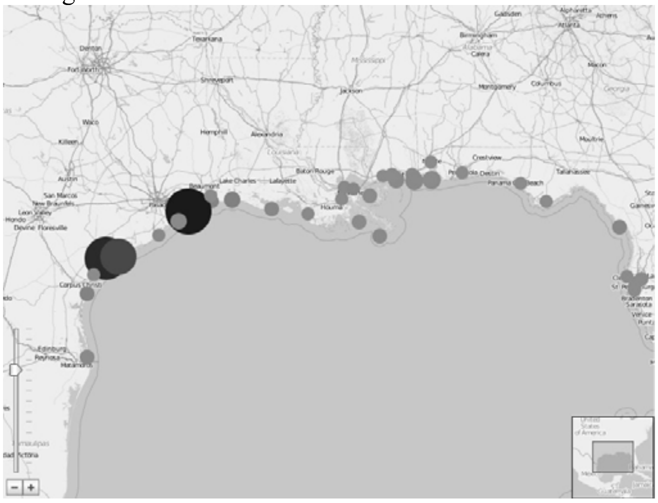

Figure 1: Good/bad hubness of sensors shown on a part of the coastline, based on wind measurements.

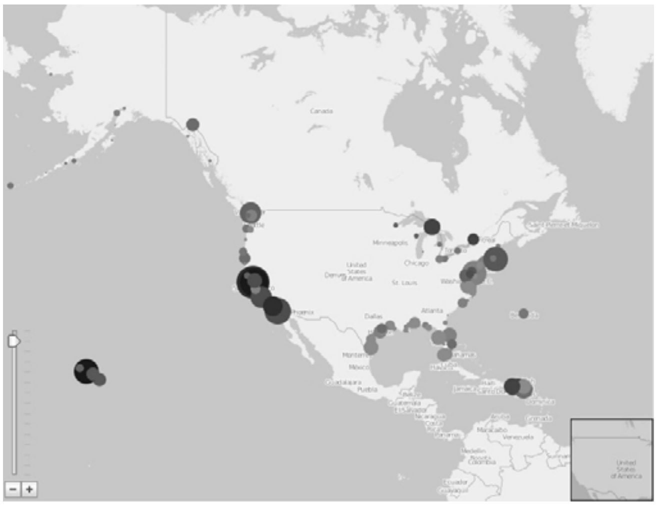

Figure 2: Good/bad hubness of sensors, based on water temperature.

By examining the results in Table 2 we see that the hubness-aware methods prove beneficial precisely in the two outlined high hubness and high bad hubness cases: wind observation and water temperature. HIKNN algorithm achieves the best result in 10 out of 12 cases with the best overall average. We see that hw-*k*NN algorithm did not achieve the best accuracy on any of the data sets. Of course, one has to have in mind that this classification setup is slightly artificial, since we do not really need to classify a sensor into a region – we know that beforehand. These tests were performed in other to test the capabilities of the listed algorithms on this type of measurement data. The results suggest that hubness-aware classification can indeed rectify the misclassifications which occur as a consequence of the underlying bad hubness present in the data.

Sensor type	***k*NN**	**hw*k*NN**	**NHBNN**	**h-FNN**	**HIKNN**
air temp.	96.8	96.7	96.0	**97.1**	96.9
bar. press.	96.8	97.0	97.0	97.0	**97.1**
wind	75.2	83.6 •	**86.0 •**	84.1 •	83.2 •
wat. l. o.	92.6	91.4	90.8	91.8	**93.3**
wat. l. p.	93.3	93.1	92.8	93.7	**94.5**
wat. tmp.	78.6	80.7	81.9	82.0	**83.3 •**
air temp.	96.2	96.0	94.0	95.9	**96.2**
bar. press.	96.9	96.5	97.1	97.2	**97.3**
wind	70.6	81.5 •	81.3 •	**82.0 •**	**82.0 •**
wat. l. o.	91.8	91.2	90.4	92.2	**92.8**
wat. l. p.	90.2	90.9	89.6	91.0	**91.7**
wat. tmp.	77.9	79.2	77.3	80.3	**82.6 •**
AVG	88.1	89.8	89.5	90.4	**90.9**

Table 2: *Classification accuracy of kNN, hw-kNN, NHBNN, h-FNN and HIKNN on sensor measurements. The upper half of the table corresponds to k=3, the lower to k=5. The best result in each line is given in bold and the statistically significant results (p < 0.05) are marked by* **•**.

5 CONCLUSION

We have explored some basic *hubness*-related properties of sensor data measured by the Integrated Ocean Observing System. Most of this data was found to exhibit low-to medium hubness, but the wind observations and water temperature measurements were more prone to the emergence of hubs. *Bad hubness* was also present in this data. By visualizing the localization of these bad hubs, it was possible to see that some of the bad hubness might actually be a consequence of erroneous data. Bad sensor hubs of the different measured properties were located at different nodes, in different regions.

Several hubness-aware classification methods: hw-*k*NN, NHBNN, h-FNN and HIKNN were tested on this data and compared to the basic *k*NN method. An improvement in accuracy was observed on the sensor types exhibiting bad hubness. HIKNN seems to be the most promising approach.

6 ACKNOWLEDGEMENTS

This work was supported by the Slovenian Research Agency, the ICT Programme of the EC under PASCAL2 (IST-NoE-216886) and PlanetData (ICT-NoE-257641).

References

[1] E.Fix and J.Hodges. Discriminatory analysis, nonparametric discrimination: consistency properties, Technical report, USAF School of Aviation Medicine, Randolph Field. Texas. 1951.

[2] T.M. Cover and P.E. Hart. Nearest neighbor pattern classification. IEEE Transactions on Information Theory. vol. IT-13. no 1. pp. 21-27. 1967.

[3] L. Devroye, L. Gyorfi and G. Lugosi. On the strong universal consistency on nearest neighbor regression function estimates. Annals of Statistics. pp. 1371-1385. 1994.

[4] K.Q. Weinberger and J. Blitzer and L.K Saul. Distance metric learning for large margin nearest-neighbor classification. *Proceedings of the NIPS conference.* MIT Press. 2006.

[5] M. Radovanović and A. Nanopulous. Nearest-neighbors in high-dimensional data: the emergence and influence of hubs. *Proceedings of 26th International Conference on Machine Learning (ICML)* pp. 865-872. 2009.

[6] N. Tomašev and M. Radovanović and D. Mladenić and M. Ivanović. Hubness-based fuzzy measures for high-dimensional k-nearest-neighbor classification. *In Proc. MDLM 2011, 7th International Conf. on Machine Learning and Data Mining*. New York. 2011.

[7] J.E. Keller and M.R. Gray and J.A. Givens. A fuzzy k nearest-neighbor algorithm. In: IEEE Transactions on Systems, Man and Cybernetics. pp. 580–585. 1985.

[8] N. Tomašev and M. Radovanović and D. Mladenić and M. Ivanović. A Probabilistic approach to nearest-neighbor classification: Naive Hubness-Bayesian kNN. *In Proc. CIKM*. 2011.

[9] N. Tomašev and D. Mladenić. Nearest-neighbor voting in high dimensional data: learning from past occurrences. (under review) 2011.

[10] N. Tomašev and M.Radovanović and D. Mladenić and M. Ivanović. The Role of hubness in clustering high-dimensional data. *In Proc. of PAKDD*. Shenzhen. 2011.

[11] M.Radovanović, A. Nanopulous and M. Ivanović. Time Series Classification in Many Intrinsic Dimensions. *In Proc. SDM.* pp. 677-688. 2010.

[12] H. Ding, G. Trajcevski, P. Scheuermann, X.Wang and E.J. Keogh. Querying and mining of time series data: Experimental comparisons of representations and distance measures. *In Proc. 34th Int. Conf. on Very Large Databases (VLDB).* pp. 1542-1552. 2008.