The Role Of Hubness in High-dimensional Data Analysis

Nenad Tomašev
Artificial Intelligence Laboratory, Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
E-mail: nenad.tomasev@gmail.com

Thesis Summary

Keywords: machine learning, curse of dimensionality, hubness

Received: November 9, 2014

This article presents a summary of the doctoral dissertation of the author, which addresses the task of machine learning under hubness in intrinsically high-dimensional data.

Povzetek: Prispevek predstavlja povzetek doktorske disertacije avtorja, ki obravnava naloge strojnega učenja pri zvezdiščnosti visokodimenzionalnih podatkov.

1 Introduction

Machine learning in intrinsically high-dimensional data is known to be challenging and this is usually referred to as the curse of dimensionality. Designing machine learning methods that perform well in many dimensions is critical, since high-dimensional data arises often in practical applications and typical examples include textual, image and multimedia feature representations, as well as time series and biomedical data.

The hubness phenomenon [1] has recently come into focus as an important aspect of the curse of dimensionality that affects many instance-based machine learning systems. With increasing dimensionality, the distribution of instance relevance within the models tends to become long-tailed. A small number of hub points dominates the analysis and influences a disproportionate number of system predictions. Most remaining points are rarely or never retrieved in relevance queries, resulting in an information loss. High data hubness has been linked to poor system performance in many data domains.

The dissertation [2] proposes several novel hubness-aware machine learning algorithms to improve the effectiveness of machine learning in intrinsically high-dimensional data. The proposed methods are based on modeling the influence of hub points on the training data.

The article is organized as follows. Section 2 gives an overview of the proposed hubness-aware methods. Section 3 summarizes the evaluation results. The dissertation's scientific contributions are outlined in Section 4 together with plans for future work.

2 Hubness-aware machine learning

2.1 Classification

The earlier approach of using hubness-based weighting in k-nearest neighbor classification was extended and several hubness-aware classification algorithms based on class-conditional k-nearest neighbor occurrence models were proposed: h-FNN, dwh-FNN, HIKNN and NHBNN. The fuzzy hubness-aware k-nearest neighbor methods (h-FNN, dwh-FNN) utilize hubness-based fuzzy measures for voting in an extended fuzzy k-nearest neighbor (FNN) framework. HIKNN further extends dwh-FNN by taking the neighbor occurrence self-information into account and assigning higher relevance to less frequently occurring neighbor points. HIKNN also removes the need for special anti-hub handling mechanisms and reduces the number of parameters needed for dwh-FNN. Unlike the fuzzy approaches, the naive hubness-Bayesian k-nearest neighbor method (NHBNN) presents a Bayesian re-interpretation of the neighbor occurrence events and offers a novel probabilistic framework for kNN classification.

2.2 Clustering

It was demonstrated in the dissertation that the neighbor occurrence frequencies in intrinsically high-dimensional data tend to be correlated with local cluster centrality. This was used to propose several deterministic and stochastic extensions of the well-known K-means clustering framework: local K-hubs (LKH), global K-hubs (GKH), local hubness-proportional clustering (LHPC), global hubness-proportional clustering (GHPC) and global hubness-proportional K-means (GHPKM) [3]. The novelty of the methods is that they exploit hubs as cluster prototypes and use point-wise hubness for guiding the search for the optimal configuration.

2.3 Metric learning

A hubness-aware extension of the commonly used $simcos_s$ shared-neighbor secondary similarity measure was proposed, $simhub_s$. The proposed approach uses hubness-based weights when calculating the secondary similarity

scores.

3 Evaluation

The proposed classification, clustering and metric learning approaches were evaluated on many intrinsically high-dimensional datasets from various domains, as well as specially generated challenging synthetic datasets. They were compared with standard hubness non-aware baselines and statistically significant improvements were observed.

The proposed approaches were shown to improve system performance in several highly challenging tasks, including learning under class imbalance and learning with feature or label noise. A disproportionate amount of misclassification in high-dimensional class-imbalanced data was determined to be caused by minority class points. Hubness-aware classifiers were found to be well suited for classification under this newly discovered *curse of minority hubs* [4].

4 Conclusions

The dissertation addresses the problem of designing effective machine learning approaches under the assumption of hubness in intrinsically high-dimensional data. It proposes several novel hubness-aware methods for learning in many dimensions. The main contributions of the dissertation are:

- Novel hubness-aware kNN classification methods: h-FNN, dwh-FNN, HIKNN and NHBNN. The novelty lies in using class-conditional neighbor occurrences for hub modeling on the training data.
- Novel clustering approaches: LKH, GKH, LHPC, GHPC, GHPKM. These are the first hubness-based clustering approaches to be proposed.
- A novel secondary similarity measure, $simhub_s$. It is the first shared-neighbor similarity score to take hubness explicitly into account.

The proposed hubness-aware approaches achieved statistically significant improvements over the corresponding baselines in various tested experimental contexts.

As future work, we wish to further extend the proposed hubness-aware approaches. We intend to pay special attention to unsupervised and semi-supervised learning, as obtaining ground truth is often expensive in practical applications. We also intend to work on scalability in order to make the proposed approaches useful on large-scale datasets.

References

[1] M. Radovanović (2011) *Representations and Metrics in High-Dimensional Data Mining*, Izdavačka knjižarnica Zorana Stojanovića, Novi Sad, Serbia.

[2] N. Tomašev (2013) *The Role of Hubness in High-dimensional Data Analysis*, IPS Jož Stefan, Ljubljana, Slovenia.

[3] N. Tomašev et al. (2013) The Role of Hubness in Clustering High-dimensional data, *IEEE Transactions on Knowledge and Data Engineering*, IEEE Computer Society, pp. 1041–4347.

[4] N. Tomašev et al. (2013) Class Imbalance and The Curse of Minority Hubs, *Knowledge-Based Systems*, Elsevier, pp. 157–172.