Knowledge-Based Systems 53 (2013) 157–172

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Class imbalance and the curse of minority hubs

Nenad Tomašev *, Dunja Mladenić

Institute Jožef Stefan, Artificial Intelligence Laboratory, Jamova 39, 1000 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 6 May 2013
Received in revised form 28 August 2013
Accepted 29 August 2013
Available online 11 September 2013

Keywords:
Class imbalance
Class overlap
Classification
k-Nearest neighbor
Hubness
Curse of dimensionality

ABSTRACT

Most machine learning tasks involve learning from high-dimensional data, which is often quite difficult to handle. *Hubness* is an aspect of the *curse of dimensionality* that was shown to be highly detrimental to *k*-nearest neighbor methods in high-dimensional feature spaces. *Hubs*, very frequent nearest neighbors, emerge as centers of influence within the data and often act as semantic singularities. This paper deals with evaluating the impact of hubness on learning under class imbalance with *k*-nearest neighbor methods. Our results suggest that, contrary to the common belief, minority class hubs might be responsible for most misclassification in many high-dimensional datasets. The standard approaches to learning under class imbalance usually clearly favor the instances of the minority class and are not well suited for handling such highly detrimental minority points. In our experiments, we have evaluated several state-of-the-art hubness-aware *k*NN classifiers that are based on learning from the neighbor occurrence models calculated from the training data. The experiments included learning under severe class imbalance, class overlap and mislabeling and the results suggest that the hubness-aware methods usually achieve promising results on the examined high-dimensional datasets. The improvements seem to be most pronounced when handling the difficult point types: borderline points, rare points and outliers. On most examined datasets, the hubness-aware approaches improve the classification precision of the minority classes and the recall of the majority class, which helps with reducing the negative impact of minority hubs. We argue that it might prove beneficial to combine the extensible hubness-aware voting frameworks with the existing class imbalanced *k*NN classifiers, in order to properly handle class imbalanced data in high-dimensional feature spaces.

1. Introduction

Nearest-neighbor methods form an important group of techniques involved in solving various types of machine learning tasks. They are based on a simple assumption that neighboring points share certain common properties. Often enough, they also share the same label, which is why so many different *k*-nearest neighbor classification algorithms have been developed over the years [28,54,36,64,53,90].

The basic *k*-nearest neighbor algorithm (*k*NN) [19] is quite simple. The label in the point of interest is derived from its *k*-nearest neighbors by a majority vote. The *k*NN rule has some favorable asymptotic properties [11].

Under the basic *k*NN approach, no model is generated in the training phase and the target function is inferred locally when the query is made to the system. Methods with this property are said to perform *lazy learning*.

Algorithms which induce classification models usually adopt the maximum generality bias [33]. In contrast, the *k*-nearest neighbor classifier exhibits high specificity bias, since it retains all the examples. The specificity bias is considered a desired property of algorithms designed for handling highly imbalanced data. Not surprisingly, *k*NN has been advocated as one way of handling such imbalanced data sets [84,33].

Data sets with significant class imbalance often pose difficulties for learning algorithms [87], especially those with a high generality bias. Such algorithms tend to over-generalize on the majority class, which in turn leads to a lower performance on the minority class. Designing good methods capable of coping with highly imbalanced data still remains a daunting task.

Certain concerns have recently been raised about the applicability of the basic *k*NN approach in imbalanced scenarios [23]. The method requires high densities to deliver good probability estimates. These densities are often closely related to class size, which makes *k*NN somewhat sensitive to the imbalance level. The difference among the densities between the classes becomes critical in the overlap regions. Data points from the denser class (usually the *majority class*) are often encountered as neighbors of points from the less dense category (usually the *minority class*). In high-dimensional

* Corresponding author. Tel.: +386 30380238.
E-mail addresses: nenad.tomasev@ijs.si (N. Tomašev), dunja.mladenic@ijs.si (D. Mladenić).

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.08.031

data the task is additionally complicated by the well known *curse of dimensionality*.

High dimensionality often exhibits a detrimental influence on classification, since all data is sparse and density estimates tend to become less meaningful. It also gives rise to the phenomenon of *hubness* [59], which greatly affects nearest neighbor methods in high-dimensional data. The distribution of neighbor occurrences becomes skewed to the right and most points either never occur in k-neighbor sets or occur very rarely. A small number of points, *hubs*, account for most of the observed neighbor occurrences. Hubs are very frequent nearest neighbors[1] and, as such, exhibit a substantial influence on subsequent reasoning.

The hubness issue first emerged in music retrieval and recommendation systems, where some songs were being too frequently retrieved, even in such cases where it was impossible to discern some reasonable semantic correlation to the queries [3,2]. Such song hubs were detrimental to the system performance. It was initially thought that this was merely a consequence of the discrepancies between the perceptual similarity and the specific similarity measures employed by the systems. It was later demonstrated that *intrinsically* high-dimensional data with finite and well-defined means has a certain tendency for exhibiting hubness [59,51,60,61] and that changing the similarity measure can only reduce, but not entirely eliminate the problem. Boundary-less high-dimensional data does not necessarily exhibit hubness [47], but this case does not arise often in practical applications. The phenomenon of hubness will be discussed in more detail in Section 3.

The fact that neighbor occurrence distributions assume a certain shape in high-dimensional data gives us additional information which can be taken into account in algorithm design. Several simple *hubness-aware* kNN classification methods have recently been proposed in an attempt to tackle this problem explicitly. An instance-weighting scheme was first proposed in [59], which reduces the bad influence of hubs during voting. An extension of the fuzzy k-nearest neighbor framework was shown to be somewhat better on average [81], introducing the concept of *class-conditional hubness* of neighbor points and building an occurrence model which is used in classification. This approach was further improved by considering the information content of each neighbor occurrence [75]. An alternative approach in treating each occurrence as a random event was explored in [79], where it was shown that some form of Bayesian reasoning might be yet another feasible way of dealing with changes in the occurrence distribution. More details on the algorithms will be given in Section 3.4.

1.1. Project goal

The phenomenon of hubness has not been studied under the assumption of class imbalance in high-dimensional data and its impact on learning with kNN methods in skewed label distributions was unknown. This raises some concerns, as most real-world data is intrinsically high-dimensional and many important problems are also class-imbalanced.

The goal of this project was to examine the influence of hubness on learning under class imbalance, as well as test the performance and robustness of the existing hubness-aware kNN classification methods in order to evaluate whether they might be appropriate for handling such highly complex classification tasks.

Most misclassification is known to occur in borderline regions, where different classes meet and overlap. Class imbalance poses a problem only if a significant class overlap is present [56], so both of these factors must be considered carefully. In our experiments, we have generated several synthetic imbalanced high-dimensional data sets with severe overlap between different distributions in order to see if the hubness-aware algorithms are able to overcome this obstacle by relying on their occurrence models.

Real-world data labels are not always very reliable. Data is usually labeled by people and people make mistakes. This is why we decided to examine the influence of very high levels of artificially induced mislabeling on the classification process.

1.2. Contributions

This research is the first attempt to correlate hubness as an aspect of the dimensionality curse with the problem of learning under class imbalance. Our analysis shows some surprising results, as our tests suggest that the minority class induces high misclassification of the majority class in many high-dimensional datasets, contrary to the low-dimensional case. We do not imply that this would always be the case, but it is an entirely new possibility that has so far been overlooked in algorithm design and needs to be carefully considered and taken into account.

We have performed an extensive experimental evaluation and shown that the recently proposed hubness-aware neighbor occurrence models achieve promising performance in several difficult types of classification problems: learning under class imbalance, mislabeling and class overlap in intrinsically high-dimensional data.

Our experiments suggest that the observed improvements stem from being able to better handle the difficult point types: borderline points, rare points and outliers. Additionally, the analysis reveals that, in most cases, the hubness-aware methods improve the recall of the majority class and the precision of the minority classes. This helps in improving the classification performance in presence of minority hubs.

Based on these encouraging results and the extensibility of the hubness-aware voting frameworks, we argue that it might be beneficial to combine them with the existing techniques for class imbalanced data classification, in order to improve system performance in high-dimensional data under the assumption of hubness.

2. Related work

2.1. Class imbalanced data classification

The problem of learning from imbalanced data has recently attracted attention of both industry and academia alike. Many classification algorithms used in real-world systems and applications fail

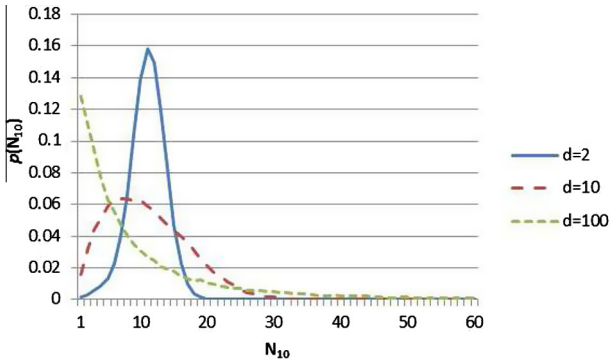

Fig. 1. The change in the distribution shape of 10-occurrences (N_{10}) in i.i.d. Gaussian data with increasing dimensionality when using the Euclidean distance. The graph was obtained by averaging over 50 randomly generated data sets. Hub-points exist also with $N_{10} > 60$, so the graph displays only a restriction of the actual data occurrence distribution.

[1] Formally, in accordance with the existing definitions in the literature [59], we will say that *hubs* are points that have an occurrence count exceeding the mean (k) by more than two standard deviations of the neighbor occurrence distribution.