Neurocomputing 160 (2015) 157–172

Contents lists available at ScienceDirect

Neurocomputing

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

Hubness-aware kNN classification of high-dimensional data in presence of label noise

Nenad Tomašev [a,*], Krisztian Buza [b]

[a] Institute Jožef Stefan, Artificial Intelligence Laboratory, Jamova 39, 1000 Ljubljana, Slovenia
[b] Institute of Genomic Medicine and Rare Disorders, Semmelweis University, Tömö utca 25-29., 1083 Budapest, Hungary

ARTICLE INFO

Article history:
Received 12 March 2014
Received in revised form
27 September 2014
Accepted 11 October 2014
Available online 11 February 2015

Keywords:
Classification
Label noise
K-nearest neighbor
High-dimensional data
Hubness
Neighbor occurrence models

ABSTRACT

Learning with label noise is an important issue in classification, since it is not always possible to obtain reliable data labels. In this paper we explore and evaluate a new approach to learning with label noise in intrinsically high-dimensional data, based on using neighbor occurrence models for hubness-aware *k*-nearest neighbor classification. Hubness is an important aspect of the curse of dimensionality that has a negative effect on many types of similarity-based learning methods. As we will show, the emergence of hubs as centers of influence in high-dimensional data affects the learning process in the presence of label noise. We evaluate the potential impact of hub-centered noise by defining a hubness-proportional random label noise model that is shown to induce a significantly higher *k*NN misclassification rate than the uniform random label noise. Real-world examples are discussed where hubness-correlated noise arises either naturally or as a consequence of an adversarial attack. Our experimental evaluation reveals that hubness-based fuzzy *k*-nearest neighbor classification and Naive Hubness-Bayesian *k*-nearest neighbor classification might be suitable for learning under label noise in intrinsically high-dimensional data, as they exhibit robustness to high levels of random label noise and hubness-proportional random label noise. The results demonstrate promising performance across several data domains.

1. Introduction

Designing effective and robust supervised learning algorithms for classification in the presence of label noise is an important practical issue, as obtaining reliable data labels is often expensive or simply infeasible due to data size in large-scale systems [18].

Classification noise can be random, feature-dependent or adversarial. Label flip probabilities can be either uniform and symmetric or depend on particular classes and class pairs. The simple *random classification noise* (RCN) model was first introduced in [2]. It is a model of how non-adversarial noise might affect the data. Given a training set $T=(X,Y)$ of labeled examples and a value $\eta \in (0, 1/2)$, $D_{\eta,T}$ denotes the distribution corresponding to T corrupted with random classification noise at rate η. A draw from $D_{\eta,T}$ is equivalent to a uniformly random draw from T where the label y of the selected (x,y) is randomly flipped with probability η.

The issue of unreliable and noisy labels can be approached in two ways: by trying to identify and correct/eliminate suspect data points and by incorporating noise into the learning model. Neither approach is trivial, as it is not always easy to distinguish mislabeled examples from the exceptions to general rules, atypical data points. When an instance lies far from its class interior and in proximity of instances from different classes, it can sometimes be mistaken for a mislabeled point [43]. Yet, atypical points sometimes hold valuable discriminative information, as they might help in defining proper class boundaries for classification. Additionally, many filtering approaches assume that all the data is available at the filtering stage and not prohibitively large [66].

Instead of filtering or explicit noise source modeling, it is also possible to design learning techniques that exhibit implicit robustness to high rates of label noise. In this paper, we will demonstrate that the recently proposed hubness-aware *k*-nearest neighbor classification methods [38,53,52,49] can be used for robust classification of intrinsically high-dimensional data under the assumption of label noise. This robustness is a consequence of the fact that, unlike in most *k*NN approaches, neighbor instances do not vote by their label at classification time. Instead, their vote is determined by their past occurrences on the training data.

Hubness [39] is a ubiquitous property of intrinsically high-dimensional data. With increasing dimensionality, the degree distribution of the *k*NN graph becomes increasingly skewed and hubs emerge as central and influential points among the data. The

* Corresponding author.
E-mail addresses: nenad.tomasev@gmail.com (N. Tomašev), chrisbuza@yahoo.com (K. Buza).

http://dx.doi.org/10.1016/j.neucom.2014.10.084
0925-2312/

*k*NN graph itself assumes a scale-free-like topology. This has multiple consequences for similarity-based learning methods and *k*-nearest neighbor methods in particular. Additionally, it changes how random labeling noise affects the learning process. Errors in hub point labels can induce severe mislabeling while errors in orphan points or regular points have little influence on the classification accuracy in *k*NN methods.

In this paper, we introduce the concept of *hubness-proportional random label noise* as an adversarial noise model where the most influential points in the data are most likely to be corrupted. The probability of a label flip is set to be proportional to the neighbor occurrence frequency of the data point. Hubness-proportional random label noise models how a potentially successful malicious attempt can compromise the most relevant and influential neighbor points in order to disrupt *k*NN-based retrieval, recommendation or prediction systems.

To our knowledge, this paper is the first detailed study dedicated to examining the influence of hubness on *k*NN classification with uncertain data labels.

The paper is organized as follows: Section 2 summarizes the related work and the existing approaches for dealing with label noise. Consequences of hubness in intrinsically high-dimensional data are discussed in Section 3. Neighbor occurrence models for hubness-aware *k*NN classification of high-dimensional data are described in detail in Section 4, followed by examples that demonstrate their potential for learning under label noise. Section 5 introduces the concept of hubness-proportional random label noise and gives practical examples of the susceptibility of *k*NN methods to label noise under high data hubness. The data used in the experiments is described in Section 6, followed by experimental results and summaries in Section 7. In Section 8, the main contributions of the paper are summarized and several directions for future work are proposed.

2. Related work

Label noise often occurs in large-scale problems where labelling is crowdsourced to a large number of non-experts instead of having the domain experts carefully label each data instance, for instance via Amazon's Mechanical Turk. In such cases, it has been shown to be beneficial to obtain multiple labels for each data point or carefully selected subsets of data points [24,42]. Evaluating the labeling accuracy of individual experts and non-experts can also be used in order to improve label quality by preferring certain labelers over others [15,60]. Modeling the concept evolution over time as the user's perception of the concepts that are being tagged by employing structured labeling has been shown to improve consistency and yield considerable improvements [30]. Unreliable labels can also result from automated information retrieval and tagging.

Data filtering for removing the mislabeled data points prior to model learning for classification is often used in practice. A simple approach is to rely on classification ensembles and to filter out those instances that are misclassified by the ensemble on the training data by taking a majority vote [9,58,65,47]. It is possible to detect data sub-samples that lead to high classification errors via cross-validation and to improve classification performance by relying on multiple data representations and discriminating subspaces [57]. Examples that lie in neighborhoods where a proportion of the dominating class is significantly lower than average are also suspect and their elimination can help with improving *k*NN classification accuracy [31]. Boolean rules inferred from the measurements can be used for detecting noisy data points [28]. Neural networks have been used for correcting the mislabeled examples in [63], by iteratively updating class affiliation probabilities based on the difference from the trained neural network output. Unlabeled examples can also be taken into account in filtering in a semi-supervised type of approach, raising the overall noise detection accuracy [22]. It is possible to formulate the noise removal task as an optimization problem, which might sometimes be preferable in comparison with the ensemble based filtering approaches [56].

Mutual information is a popular feature selection criterion and a robust estimation of mutual information via a probabilistic noise model was able to improve feature selection performance under label noise [17]. This was achieved by an adaptive hyper-sphere radius selection in nearest-neighbor entropy estimators. Certain feature extraction strategies have been successfully employed to improve classification accuracy in noisy medical data [35].

The presence of label noise in class-imbalanced learning tasks can be highly detrimental and it was shown to affect the learning process differently depending on whether mislabeling occurs in the minority or the majority class [23]. This is important as most noise removal strategies treat these two cases equally.

Non-uniform label noise sometimes arises due to systematic errors in data acquisition or the experimental design that produces the data in question [36,21]. The type of noise should be determined prior to deciding on the optimal noise handling strategy.

As individual labels are unreliable, it is possible to use multi-instance learning in order to aggregate instances and assign labels to groups of instances instead. This has been shown to be a promising approach [25].

While boosting methods may be popular in practice [11,26,65], recent research suggests that many types of boosting methods that can be interpreted as convex potential boosters are highly susceptible to random classification noise [32]. Branching program based boosters that do not fall into this framework can still achieve good learning accuracy on noisy data.

Designing classifiers that are able to implicitly handle noisy and mislabeled data points is another approach and one such classifier is the adaptive *k*-nearest neighbor classifier (AKNN) [59] that re-scales the distances of training points to the query, based on their proximity to the closest point of a different class. As labels in mislabeled points often do not match the labels among their neighbors, this approach disregards most mislabeled points as it adaptively increases their distance to the query. This approach will be our baseline for the experiments in Section 7. Deep learning algorithms can be extended to handle label noise by additional network layers for noise modeling [45]. Robust kernels can be learned from the data in order to improve the effectiveness of kernel-based methods under label noise [7] and robust SVM methods have also been considered [44,6].

The existing noise-handling strategies fail to take data hubness into account and do not attribute special attention to potential errors in the hub points, which might be an issue when learning from high-dimensional data. This problem was identified in [10], where it was noted that a surprising number of classification errors in time series *k*NN classification can be attributed to hub points.

3. Hubness in intrinsically high-dimensional data

Hubness is a consequence of high intrinsic data dimensionality related to the degree distribution of the *k*NN graph [39]. Hub points arise as centers of influence, as they occur very frequently as nearest neighbors. In fact, the entire neighbor occurrence frequency distribution becomes skewed and most points become *anti-hubs* or *orphans*, i.e. they occur rarely or never as neighbors to other points. Hubs often exhibit a detrimental influence by