Knowledge-Based Systems xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Nearest neighbor regression in the presence of bad hubs

Krisztian Buza [a,*], Alexandros Nanopoulos [b], Gábor Nagy [c]

[a] BioIntelligence Lab, Genomic Medicine and Rare Disorders, Semmelweis University, Hungary
[b] University of Eichstätt-Ingolstadt, Germany
[c] Budapest University of Technology and Economics, Hungary

ARTICLE INFO

Article history:
Received 1 December 2014
Received in revised form 8 June 2015
Accepted 9 June 2015
Available online xxxx

Keywords:
Nearest neighbor regression
Hubs

ABSTRACT

Prediction on a numeric scale, i.e., regression, is one of the most prominent machine learning tasks with various applications in finance, medicine, social and natural sciences. Due to its simplicity, theoretical performance guarantees and successful real-world applications, one of the most popular regression techniques is the *k* nearest neighbor regression. However, *k* nearest neighbor approaches are affected by the presence of *bad hubs*, a recently observed phenomenon according to which some of the instances are similar to surprisingly many other instances and have a detrimental effect on the overall prediction performance. This paper is the first to study bad hubs in context of regression. We propose hubness-aware nearest neighbor regression schemes. We evaluate our approaches on publicly available real-world datasets from various domains. Our results show that the proposed approaches outperform various other regressions schemes such as *k*NN regression, regression trees and neural networks. We also evaluate the proposed approaches in the presence of label noise because tolerance to noise is one of the most relevant aspects from the point of view of real-world applications. In particular, we perform experiments under the assumption of conventional Gaussian label noise and an adapted version of the recently proposed hubness-proportional random label noise.

1. Introduction

Regression, i.e., prediction of a continuous target variable from a set of observations, is one of the most prominent machine learning tasks with various applications in engineering, finance, industry and medicine, see e.g. [1–5]. Various regression techniques have been developed in the last decades ranging from simple linear and polynomial regression to more complex models, such as neural networks [6,7] and support vector regression [8].

In many cases, not only the nominal dimensionality of the data is high, but the same is true for the number of *meaningful* (or *intrinsic*) dimensions, although the later may vary from instance to instance: for example, the medical records of a patient p may involve the results of different examinations than the records of another patient p' (who may suffer from different diseases than p). Therefore, making use of such data is not only difficult because of its size, but also due to its complexity: without loss of essential information, calculating the distance or similarity between instances may already be rather challenging, while finding a reasonable vector representation of the data may be even more difficult. On the other hand, there is an ever-growing interest in using such semi-structured data for prediction, which involves prediction on a numeric scale, i.e., regression. Consequently, in this paper we focus on regression techniques that only assume the presence of an appropriate distance or similarity measure which *may* or *may not* be based on the vector representation of the instances.

Despite the aforementioned variety of regression schemes, one of the most popular techniques is the nearest neighbor regression. While being intuitive, nearest neighbor regression is well-understood from the point of view of theory, see e.g. [9–11] and the references therein for an overview of the most important theoretical results regarding the performance of nearest neighbor regression. These theoretical results are also justified by empirical studies: for example, in their recent paper, Stensbo-Smidt et al. found that nearest neighbor regression outperforms model-based prediction of star formation rates [12], while Hu et al. showed that a *k*-nearest neighbor regression based model is able to estimate the capacity of lithium-ion batteries [3].

We point out that most of the conventional regression schemes were developed for vector data, i.e., under the assumption that the data can be organized into a data table with well-defined dimensionality, whereas in the aforementioned cases this assumption

* Corresponding author.
E-mail addresses: buza@biointelligence.hu (K. Buza), alexandros.nanopoulos@ku.de (A. Nanopoulos), nagy.gabor.i@gmail.com (G. Nagy).

http://dx.doi.org/10.1016/j.knosys.2015.06.010
0950-7051/

may be violated. However, nearest neighbor-models only require a distance or similarity between the instances, which may be much simpler to define than finding a suitable vector representation of the data, see e.g. edit distances for time series, genetic sequences or texts, such as dynamic time warping [13,14], Smith–Watermann distance [15] or Levenshtein distance [16]. These distance measures work directly on the "raw" data (i.e., time series, genetic sequences or texts respectively) without an intermediate vector representation.

Machine learning in high dimensional data spaces is particularly challenging due to the phenomena known under the umbrella of the *curse of dimensionality*. One of the recently explored aspects of the curse is the emergence of bad hubs, see e.g. [17–21]. Informally, hubs are instances that are similar to a surprisingly high amount of other instances. Unfortunately, some of the hubs are bad in the sort of sense that they may mislead classification algorithms. While bad hubs are well-studied in case of classification [22], instance selection [23] and clustering [24], in context of regression problems bad hubs have not been described yet. Providing an analysis of the presence of bad hubs in regression problems is not trivial because the original definition of bad hubs assumes discrete class labels, however, in case of regression problems, the labels are continuous. Therefore, in order to study bad hubs in context of regression, we need a novel approach.

In this paper, we focus on nearest neighbor regression and study the presence of bad hubs in context of regression problems. Motivated by these observations, we propose hubness-aware nearest neighbor regression schemes. Subsequently, we evaluate our approach on publicly available real-world datasets from various domains: prediction of yields on the stock market, assessment of the severity of Parkinson's disease, estimation of the area of forest fires, prediction of the number of comments that a blog post will receive and assessment of wine quality. Our experimental results show that our approach is favorable in all these domains. Additionally, we evaluate the proposed approaches in the presence of label noise because, on the one hand, tolerance to noise is one of the most relevant aspects from the point of view of real-world applications, on the other hand, the selection of appropriate noise models allow us to simulate the increased presence of bad hubs. In particular, we perform experiments under the assumption of two types of noise: we consider conventional Gaussian label noise and an adapted version of the recently proposed hubness-proportional random label noise [25]. This adaptation is one of the minor contributions of the paper and it is necessary because hubness-proportional random label noise was originally introduced for classification problems. According to the best of our knowledge, this is the first paper that studies the presence of bad hubs in context of regression problems, and this is the first paper that proposes hubness-aware regression schemes and evaluates them both on real-world datasets and under various noise models.

2. Definitions and notations

A dataset $\mathcal{D}$ containing n instances is given. Instances are denoted by $x_i, 1 \leqslant i \leqslant n$. For each instance $x_i \in \mathcal{D}$, the value of the continuous target is given and it is denoted by $y(x_i)$. We say that $y(x_i)$ is the *label* of instance x_i and $\mathcal{D}$ is the training dataset. With regression we mean the task of predicting (estimating) the label of an instance $x' \notin \mathcal{D}$.

We propose a regression approach that is independent of the representation of the instances, the only requirement is that distances can be defined between the instances. Therefore, we use $d(x_i, x_j)$ to denote the distance between two instances x_i and x_j.

Assume that we want to predict the label of an instance $x' \notin \mathcal{D}$. Nearest neighbor regression determines the k nearest neighbors of x', i.e., a subset $\mathcal{N}_k^{\mathcal{D}}(x')$ of $\mathcal{D}$ so that

$$\left|\mathcal{N}_k^{\mathcal{D}}(x')\right| = k \tag{1}$$

and

$$\max_{x \in \mathcal{N}_k^{\mathcal{D}}(x')} d(x', x) \leqslant \min_{x \in \mathcal{D} \setminus \mathcal{N}_k^{\mathcal{D}}(x')} d(x', x). \tag{2}$$

We may omit the upper index $\mathcal{D}$, whenever it is unambiguous in which dataset we search for the nearest neighbors of x'. Nearest neighbor regression [26,27] estimates the value of the target as the average of the labels of the nearest neighbors:

$$\hat{y}(x') = \frac{1}{k} \sum_{x_j \in \mathcal{N}_k(x')} y(x_j). \tag{3}$$

3. Bad hubs in regression problems

We note that the k nearest neighbor relationship is asymmetric: while each instance $x \in \mathcal{D}$ has k nearest neighbors, an instance $x' \in \mathcal{D}$ does not necessarily appear k-times as one of the k nearest neighbors of other instances. This is illustrated in Fig. 1 for $k = 1$. In order to keep the example simple, we consider two-dimensional vector data, therefore, instances correspond to points of the plane. In Fig. 1, instances are denoted by circles. There is a directed edge from each instance to its first nearest neighbor. While each instance has *exactly one* first nearest neighbor, i.e., the number of *outgoing* edges is exactly one for each instance; how many times an instance appears as the first nearest neighbor of *other* instances, i.e., the number of *incoming* edges, is not necessarily one. As one can see, some of the instances never appear as nearest neighbors of others and there is an instance that appears as the first nearest neighbor of three other instances: the integer number next to each instance shows how many times it appears as the first nearest neighbor of others.

Generally, we use $N_k(x)$ to denote how many times the instance $x \in \mathcal{D}$ appears as one of the k nearest neighbors of other instances of $\mathcal{D}$. It is easy to see that the expected value of $N_k(x)$ is $E[N_k(x)] = k$, however, the actual value of $N_k(x)$ varies from instance to instance. While considering k nearest neighbor models, $N_k(x)$ can be seen as the measure of how influential is the instance x. As it was shown in previous works, see e.g. [17,18,23], in many cases, the distribution of $N_k(x)$ is substantially skewed to the right, i.e., there are a few instances with extraordinarily high $N_k(x)$ values. Usually, instances having surprisingly high $N_k(x)$ are called *hubs*, while instance with exceptionally low $N_k(x)$ are called *anti-hubs*. More precisely, we say that an instance x is a hub, if $N_k(x) > 2k$; while an instance x is an *anti-hub* if $N_k(x) = 0$. The phenomenon that $N_k(x)$ is skewed is called *hubness* and it is often quantified by the third standardized moment (skewness) of the distribution of $N_k(x)$.

In case of classification, we say that an instance x is a *bad k nearest neighbor* of another instance x' if x is one of the k-nearest neighbors of x' and the both instances have different class labels. Consequently, in case of classification, *bad k-occurrence* $BN_k(x)$ of an instance x was introduced to measure how many times an instance x appears as bad nearest neighbor of other instances. Similarly to the distribution of $N_k(x)$, the distribution of $BN_k(x)$ was shown to be substantially skewed in case of high-dimensional data.

Similar observations can be made for high-dimensional data associated with numerical prediction tasks. As an example, in the left of Fig. 2 we show the distribution of $N_k(x)$ for the Financial Tweets Data using $k = 10$ and the Euclidean distance. The dataset

Please cite this article in press as: K. Buza et al., Nearest neighbor regression in the presence of bad hubs, Knowl. Based Syst. (2015), http://dx.doi.org/10.1016/j.knosys.2015.06.010