CSIAM Trans. Appl. Math.
doi: 10.4208/csiam-am.2020-0212

Vol. **2**, No. 1, pp. 81-107
March 2021

Identification of Corrupted Data via k-Means Clustering for Function Approximation

Jun Hou[1], Yeonjong Shin[2] and Dongbin Xiu[1,*]

[1] *Department of Mathematics, The Ohio State University, Columbus, OH 43210, USA.*
[2] *Division of Applied Mathematics, Brown University, Providence, RI 02912, USA.*

Received 2 September 2020; Accepted 24 November 2020

Abstract. In addition to measurement noises, real world data are often corrupted by unexpected internal or external errors. Corruption errors can be much larger than the standard noises and negatively affect data processing results. In this paper, we propose a method of identifying corrupted data in the context of function approximation. The method is a two-step procedure consisting of approximation stage and identification stage. In the approximation stage, we conduct straightforward function approximation to the entire data set for preliminary processing. In the identification stage, a clustering algorithm is applied to the processed data to identify the potentially corrupted data entries. In particular, we found k-means clustering algorithm to be highly effective. Our theoretical analysis reveal that under sufficient conditions the proposed method can exactly identify all corrupted data entries. Numerous examples are provided to verify our theoretical findings and demonstrate the effectiveness of the method.

AMS subject classifications: 42C05, 41A10, 65D15

Key words: Data corruption, function approximation, sparse approximation, k-means clustering.

1 Introduction

Real world data are never perfect—in addition to standard measurement noises, they are often corrupted by unexpected and uncontrollable internal or external errors. The causes of corruption include human processing errors, data transmission or storage errors, machine malfunction during data collection, etc. The resulting corrupted errors can be large in magnitude and do not follow certain statistical laws. The presence of data corruptions thus can significantly impact data analysis results in a negative manner.

*Corresponding author. *Email addresses:* hou.345@osu.edu (J. Hou), yeonjong_shin@brown.edu (Y. Shin), xiu.16@osu.edu (D. Xiu)

In this paper, we consider the problem of identifying data corruptions in the context of regression modeling (supervised learning). Our approach is motivated by *function approximation with corruptions* [8,30]. Let $f(x)$ be an unknown function defined in a bounded domain D, $x_i \in D$ be an input data and $f(x_i)$ be its corresponding clean output value for $i=1,\ldots,m$. We are interested in the case where the data vector is corrupted by unexpected external errors that may be caused by aforementioned reasons. That is, the available data vector is given by

$$y = f + e_s,$$

where $f = (f(x_1),\cdots,f(x_m))^T$ is the clean output vector (which may contain the standard noises), and $e_s \in \mathbb{R}^m$ is the corruption vector with sparsity s, which stands for the number of corrupted data entries. While the vector y is the available data vector, no information on e_s is available.

A general procedure of approximating functions can be described as a class of minimization problem. Given a set of basis $\{\phi_j\}_{j=1}^n$ in D, we consider an approximation in the form of $\widetilde{f}(x) = \sum_{j=1}^n c_j \phi_j(x)$. We are interested in the oversampled case, $m > n$. The standard approach seeks to find the coefficients $c = (c_1,\cdots,c_n)^T$ that minimize the errors, i.e.,

$$\min_c \|y - Ac\|, \quad \text{where } A = \big(a_{ij}\big) = \big(\phi_j(x_i)\big) \text{ and } y = (y_i).$$

We note that the available data y is contaminated by e_s and the clean output f is not available to us. The use of the vector 2-norm yields the well-known least squares (LSQ) method, whose literature is too large to mention here. In general, LSQ method is known to be robust when the corruption errors are relatively small (e.g. Gaussian noise). The use of the vector 1-norm yields the ℓ_1 minimization, which is called least absolute deviations (LAD) is also studied extensively in [1, 4, 6, 26, 27, 30, 31]. The LAD method is known to be robust against outliers and sparse corruptions [30]. In the spirit of seeking sparsity, one can also employ any sparse approximation techniques that include ℓ_{1-2} minimization [23, 33–35] (the difference between the 1-norm and the 2-norm), or ℓ_p minimization [10, 11, 32] for $0 < p < 1$. Although these methods are capable of producing accurate function approximation, *they can not detect corrupted data*, especially when the number s of corrupted data is unknown.

In this paper, we present an approach for identifying the corrupted data entries in a given measurement data vector without the knowledge of the number (s) of the corrupted data entries. We propose a two-step procedure that consists of approximation and identification stages. At the approximation stage, we conduct function approximation with the corrupted data and obtain a residual vector. At the identification stage, we apply a clustering algorithm to the residual vector to separate the residues into corrupted entries and clean entries. Specifically, we employ k-means clustering [3, 17, 24], a well-established clustering algorithm with a wide range of applications [5, 13, 21, 25]. We then provide theoretical results on the sufficient conditions under which the proposed approach can detect the corrupted data exactly (Theorem 4.1).