Neural Networks 59 (2014) 23–35

Contents lists available at ScienceDirect

Neural Networks

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

Model-wise and point-wise random sample consensus for robust regression and outlier detection

Moumen T. El-Melegy *

Department of Electrical Engineering, Assiut University, Assiut 71516, Egypt

ARTICLE INFO

Article history:
Received 13 February 2013
Received in revised form 20 May 2014
Accepted 23 June 2014
Available online 7 July 2014

Keywords:
Multi-layered feed-forward neural networks
Training algorithm
Robust statistics
Regression
Outliers

ABSTRACT

Popular regression techniques often suffer at the presence of data outliers. Most previous efforts to solve this problem have focused on using an estimation algorithm that minimizes a robust M-estimator based error criterion instead of the usual non-robust mean squared error. However the robustness gained from M-estimators is still low. This paper addresses robust regression and outlier detection in a random sample consensus (RANSAC) framework. It studies the classical RANSAC framework and highlights its model-wise nature for processing the data. Furthermore, it introduces for the first time a point-wise strategy of RANSAC. New estimation algorithms are developed following both the model-wise and point-wise RANSAC concepts. The proposed algorithms' theoretical robustness and breakdown points are investigated in a novel probabilistic setting. While the proposed concepts and algorithms are generic and general enough to adopt many regression machineries, the paper focuses on multilayered feed-forward neural networks in solving regression problems. The algorithms are evaluated on synthetic and real data, contaminated with high degrees of outliers, and compared to existing neural network training algorithms. Furthermore, to improve the time performance, parallel implementations of the two algorithms are developed and assessed to utilize the multiple CPU cores available on nowadays computers.

1. Introduction

Regression analysis seeks to find the relationship between one or more independent variables and a dependent variable. Quite often in order to find this relationship, a mean squared error (MSE) criterion is minimized. This criterion is sufficient (indeed optimal) to deal with data corrupted by noise that follows a Gaussian model. However in many real applications, data are not only noisy but also contain outliers, data that are in gross disagreement with a postulated model. It has been noted that the occurrence of outliers in routine data ranges from 1% to 15% (Hampel, Ronchetti, Rousseeuw, & Stahel, 1986). In applications where data are collected using fully automated techniques, it is not surprising that an outlier rate as high as 50% may be encountered (Ahmed & Farag, 2002; Meer, Mintz, Rosenfeld, & Kim, 1991; Zhang, 1998). Those inevitable outliers can severely distort a fitting process so that the estimated parameters become useless. As a matter of fact, just one bad outlier would skew the results of any approach based on the mean squared estimates.

Because of the skewing of the result, trying to detect outliers by thresholding on the residual errors will not work in general. Throwing away one datum at a time and doing mean squares on the remaining subset often does not work when more than one outlier are present. As such, the estimation problem becomes harder as it is actually two problems: classification of data into inliers (valid data) and outliers; and fitting the model to data in an optimal manner. In such circumstances the deployment of robust estimation methods is essential. Robust methods continue to recover meaningful descriptions of a statistical population even when the data contain outlying elements belonging to a different population. They are also able to perform when other assumptions underlying the estimation, say the noise model, are not wholly satisfied. Several robust techniques have been proposed in the field of robust statistics (Hampel et al., 1986; Huber, 1981; Rousseeuw & Roy, 1987). These methods include M-estimates (Maximum likelihood estimates), L-estimates (linear combination of order statistics), R-estimates (estimates based on rank transformations) and LMedS estimates (Least Median Square).

The RANdom SAmple Consensus (RANSAC) framework of Fischler and Bolles (1981) has become the standard way of dealing with outliers in the computer vision and image processing community (Haralick, 1986; Meer et al., 1991; Zhang, 1998; Zhuang, Wang, & Zhang, 1992). It is able to cope with as large as 50% of

* Tel.: +20 88 2411021.
E-mail address: moumen@aun.edu.eg.

http://dx.doi.org/10.1016/j.neunet.2014.06.010
0893-6080/

outlying data. RANSAC searches for suitable solutions directly using the data, repeatedly constructing solutions from randomly sampled minimal subsets, and then tests the solution for support from the complete set of data. In RANSAC the support is the number of data items within a distance threshold. The solution from the subset with the most support is deemed the robust outcome.

The main goal of this paper is to develop novel concepts and estimation algorithms in the RANSAC framework for the objective of robust regression and outlier detection. The paper carefully studies the classical RANSAC framework and highlights its model-wise nature for processing the data. Furthermore, it introduces for the first time a point-wise strategy of RANSAC. New algorithms are developed following both the model-wise and point-wise RANSAC concepts for estimating regression models in the presence of outlier. While the proposed concepts and algorithms are generic and general enough to adopt many regression machineries, such as regression trees and support vector machines (SVMs), this paper focuses on multilayered feed-forward neural networks (MFNNs) (Haykin, 2008; Zurada, 1992) due to their popularity in solving regression problems in many diverse and practical applications. MFNNs have demonstrated a great ability to solve outlier-free regression problems and to accurately approximate unknown functions. However they suffer at the presence of outliers.

Several robust training algorithms have been proposed for feed-forward neural networks, e.g., Chen and Jain (1994); Chuang, Su, and Hsiao (2000); El-Melegy, Essai, and Ali (2009); Liano (1996); Pernia-Espinoza, Ordieres-Mere, de Pison, and Gonzalez-Marcos (2005) and Rusiecki (2010). Almost all these methods rely on M-estimators (Hampel et al., 1986; Huber, 1981; Rousseeuw & Roy, 1987) to reduce the effect of outliers on the training outcome. Unfortunately these estimators do not possess a high enough *breakdown point*, which is the smallest proportion of outliers that may force the value of the estimate to be arbitrary wrong (Kim, Kim, Meer, Mintz, & Rosenfeld, 1989; Rousseeuw & Roy, 1987). Reported experiments (Chen & Jain, 1994; Chuang et al., 2000; Liano, 1996; Pernia-Espinoza et al., 2005) demonstrated a breakdown of 10%–15% at most. In fact, several researchers tested the robustness of their methods on synthetic data by introducing a few outliers to the experiment data. El-Melegy et al. (2009) developed a training algorithm based on the more robust Least-median-of-squares estimator (LMedS) that has *theoretically* the maximum outliers insensitivity as it can tolerate up to 50% of outlying data (Huber, 1981; Rousseeuw & Roy, 1987). Nevertheless, unlike the M-estimators, there is no straightforward method (explicit formula) to minimize the LMedS estimator. Therefore the authors (El-Melegy et al., 2009) resort to the stochastic simulated annealing (SA) algorithm to minimize this estimator. Approaches based on the adaptive learning rate (Rusiecki, 2006) and Least Trimmed Squares (LTS) (Rusiecki, 2007) estimator were also proposed. The concept of initial data analysis by the Minimum Covariance Determinant (MCD) estimator was also investigated (Rusiecki, 2008). An annealing dynamical learning algorithm combined with particle swarm optimization was also proposed to train wavelet neural networks for identifying nonlinear systems with outliers (Ko, 2012). Some robust learning algorithms have been also applied to radial basis function networks (Chuang, Jeng, & Lin, 2004; Lee, Chung, Tsai, & Chang, 1999).

In spite of its popularity in the computer vision and image processing communities, the RANSAC framework has been less known in the neural network community. The first key contribution of the paper is the presentation of the RANSAC framework to the neural network research community for the objective of robust regression and outlier detection. We are aware of only one research effort (Nishida & Kurita, 2008) that used a RANSAC-like method to improve the performance of SVMs on large-scale datasets by training a number of small SVMs for randomly selected subsets of data, while tuning their parameters to fit SVMs to the whole training set. However any possibility of outlying data was not considered, and the concept of robustness was not addressed at all. The current paper adapts the model-wise RANSAC framework, for the first time, for the robust training of MFNNs. In order to make the adaptation of this framework for this goal more efficient, several novel aspects are introduced in the standard framework. A stage following the standard RANSAC procedure is employed to improve the outlier/inlier data classification, and to increase the accuracy of the estimated model representing the inlying data. Furthermore, to enhance the sampling process of the RANSAC algorithm, a selection mechanism is proposed based on bucketing techniques (Zhang, Deriche, Faugeras, & Luong, 1995). This mechanism aims to avoid wasting computational time over useless sampled subsets. Some early draft of these ideas has appeared in our recent papers (El-Melegy, 2011a), upon which we build in this paper.

A second key contribution in this paper is the development of another new RANSAC algorithm for robust regression. This algorithm proceeds in a way similar to the first RANSAC algorithm, but follows a different strategy as it keeps track of the support for a data point to be valid, instead of computing the support for the estimated model. The consensus of random subsets is taken point-wise rather than model-wise. This new point-wise RANSAC strategy has not been used before in the entire RANSAC literature. Both proposed algorithms are extensively evaluated for nonlinear regression on synthetic and real data, contaminated with varying degrees of outliers under different scenarios, as well as real high-dimensional data publicly available from the standardized benchmark collection for neural networks PROBEN1 (Prechelt, 1994) and the UCI repository of standard machine learning datasets. Both algorithms successfully outperform well-known algorithms in the literature based on M- and LMedS-estimators. It is noteworthy that the algorithms' robustness can be further improved through integrating an M-estimator based criterion within the algorithms.

The third contribution of this paper is the theoretical investigation of the algorithms' robustness and breakdown points in a novel probabilistic setting. To the best of our knowledge, this has not been reported before in the literature for any RANSAC-based algorithm. Most previous researchers have settled for the underlying reasoning behind the first RANSAC algorithm (Fischler & Bolles, 1981) or for some experimental validations (Chum & Matas, 2008; Meer et al., 1991; Torr & Zisserman, 2000; Zhuang et al., 1992), but no explicit formulas have been given. The fourth contribution of the paper deals with the time performance of the proposed algorithms. Due to its repetitive sampling and estimation nature, any RANSAC algorithm typically takes a prolonged period of time till convergence. The paper presents parallel implementations of the two proposed RANSAC algorithms and the assessment of their time performance on a multi-core desktop computer.

The rest of this paper is organized as follows. Section 2 defines first the regression problem, and then develops a model-wise RANSAC algorithm for the robust training of MFNNs. A new point-wise RANSAC algorithm is proposed in Section 3. The robustness of the two algorithm is theoretically analyzed in Section 4. Our experimental results are described in Section 5. Parallel implementations of the RANSAC algorithms are evaluated in Section 6. We give our conclusions and future work in Section 7.

2. RANSAC training of neural networks

Let us start first by stating the regression problem under concern in this paper. Let $\mathcal{S} = \{(\mathbf{x}_i, y_i),\ i = 1, \ldots, n\}$ be a training set of input–output pairs, where $\mathbf{x}_i \in \mathbb{R}^m$ are m-dimensional input data points, and $y_i \in \mathbb{R}$ are the corresponding outputs. For some pairs $\{(\mathbf{x}_k, y_k) \in \mathcal{S} | k = 1, \ldots, n',\ n' < n\}$, the inputs and outputs are related by an unknown (nonlinear) function f such that $y_k = f(\mathbf{x}_k) + \varepsilon_k$. The errors ε_k are typically assumed independent and