Available online at www.sciencedirect.com
ScienceDirect
St. Petersburg Polytechnical University Journal: Physics and Mathematics 000 (2017) 1–8
www.elsevier.com/locate/spjpm

Clustering algorithms application to forming a representative sample in the training of a multilayer perceptron

Aleksey A. Pastukhov*, Aleksander A. Prokofiev

National Research University of Electronic Technology, 5 Pass. 4806, Zelenograd, Moscow 124498, Russian Federation

Available online xxx

Abstract

In this paper, we have considered the problem of effectively forming the representative sample for training a neural network of the multilayer perceptron (MLP) type. An approach based on the use of clustering that allowed to increase the entropy of the training set was put forward. Various clustering algorithms were examined in order to form the representative sample. The algorithm-based clustering of factor spaces of various dimensions was carried out, and a representative sample was formed. To verify our approach we synthesized the MLP neural network and trained it. The training technique was performed with the sets formed both with and without clustering. A comparative analysis of the effectiveness of clustering algorithms was carried out in relation to the problem of representative sample formation.

Keywords: Neural network; Clustering algorithm; Representative sample; Multilayer perceptron.

Introduction

Training a neural network type of the multilayer perceptron (MLP) type involves a data preprocessing stage that must be completed before the back-propagation algorithm can be applied. The majority of the studies published on the application of neural networks limited the preprocessing techniques to normalization, scaling and weight pre-initialization.

While these operations are undoubtedly necessary, they can hardly be considered sufficient. For a factor space with a small dimension, the specifics of the distribution of the initial data needs to be taken into account to effectively train the neural network. This task is substantially complicated with a large number of factors. In such a case, it is advisable to apply clustering to form a training set consisting of examples with the most unique attributes in the set.

There are numerous clustering algorithms, but all of them can be conditionally divided into two groups: crisp and fuzzy. In turn, there are two types of crisp methods: hierarchical and non-hierarchical ones [1].

A separate class includes clustering algorithms based on neural networks, which have found wide application in various fields.

For example, Ref. [2] explored data clustering based on a Markov algorithm and based on self-organizing growing neural networks. Ref. [3] compared the *k*-means clustering and density-based

* Corresponding author.
E-mail addresses: pastukhov.alexei@gmail.com, pastuhov1992@gmail.com (A.A. Pastukhov), aaprokof@yandex.ru (A.A. Prokofiev).

http://dx.doi.org/10.1016/j.spjpm.2017.05.004

(Peer review under responsibility of St. Petersburg Polytechnic University).

Please cite this article as: A.A. Pastukhov, A.A. Prokofiev, Clustering algorithms application to forming a representative sample in the training of a multilayer perceptron, St. Petersburg Polytechnical University Journal: Physics and Mathematics (2017), http://dx.doi.org/10.1016/j.spjpm.2017.05.004

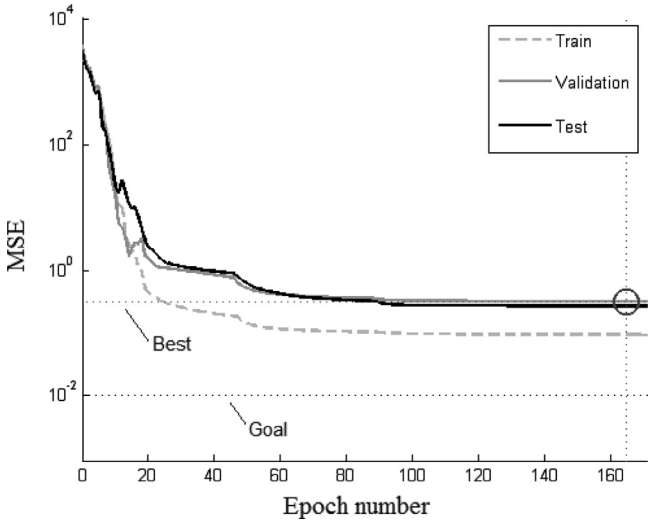

Fig. 1. Training results for the MLP neural network from the data formed without clustering: MSE is the mean square training error; Epoch number is the current epoch. The error patterns are shown for the training, validation and test sets; the goal and best error values are marked; the latter was achieved for the test set.

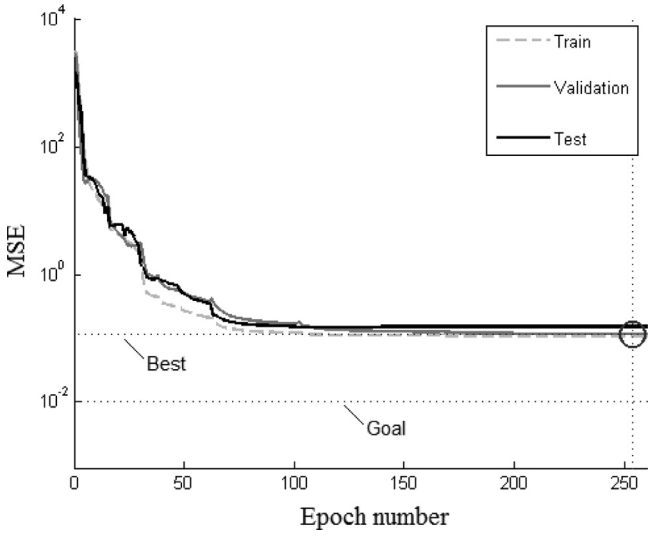

Fig. 2. Results of the MLP neural network training on the data generated using the SOM algorithm.

clustering of applications with noise (DBSCAN) in a random sample, and assessed the efficiency of these algorithms based on the Davis–Bouldin index. Ref. [4] is dedicated to clustering of text documents for creating automated classification systems using the Euclidean–Mahalanobis distance, and Ref. [5] discusses applying various algorithms of neural associative memory for creating the memory of an anthropomorphic robot.

In this paper, we have examined one fuzzy and three crisp clustering methods within the problem of forming representative samples for training a neural network; we have analyzed the efficiency of these algorithms from the standpoint of increasing the entropy of the training set and improving the training quality for an MLP-type neural network. We have additionally analyzed the variations in the entropy of the training set and in the mean square error (MSE) of training via these algorithms.

The crisp clustering algorithms discussed in this paper are the *k*-means algorithm [3] that is the most common and easiest to implement, the self-organizing Kohonen maps [6] (the maps are considered in Ref. [7]), and a clustering algorithm based on constructing a hierarchical tree of clusters [8].

Please cite this article as: A.A. Pastukhov, A.A. Prokofiev, Clustering algorithms application to forming a representative sample in the training of a multilayer perceptron, St. Petersburg Polytechnical University Journal: Physics and Mathematics (2017), http://dx.doi.org/10.1016/j.spjpm.2017.05.004