Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 112 (2017) 2363–2371

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France

Local regression algorithms based on centroid clustering methods

Marcin Pietrzykowski[a]

[a]Faculty of Computer Science and Information Technology, West Pomeranian University of Technology, Żołnierska 49, 71-210 Szczecin, Poland

Abstract

The paper analyzes different version of the mini-model method (MM-method) which are based on centroid clustering algorithms. The article introduces two new versions of the method which are based on *k*-medoids method and the fuzzy clustering methods - *c*-means algorithm. The MM-method is an instance-based learning algorithm. It operates only on data from the local neighborhood of a query. The local neighborhood is created by clustering algorithms. It makes the learning procedure simpler, because in the previous version, the neighborhood of query was based on multidimensional polytopes. The article also presents comparison between different versions of the MM-method and other instance-based learning algorithms.

Peer-review under responsibility of KES International

Keywords: mini-model, local self-learning, function approximation, instance-based learning, *k*-means, *k*-medoids, *c*-means

1. Introduction

Mini-model (MM-method) is a method of local modeling. At the beginning it should be explained that there exists a meaningful difference between a global model and a mini-model (which is local and space-constrained). For example in 2D-space both of the model-types can use the same method of approximation - linear regression. However, the global model is identified on the basis of all available data, whilst the mini-model uses only a part of samples from the entire data set. Data samples on which mini-model is learned, belong to the space-constrained local neighborhood of a query point. That neighborhood is called mini-model domain (MM-domain) or a mini-model area and is defined only in the input space of the problem. In an extreme case the MM-domain can be extended to the whole system domain, but this situation is very rare and possible only when a modeled data fits to the mini-model in an entire domain.

The concept of mini-models method was developed by Piegat[1,2,3]. MM-method depends on assumption that in the modeling task very often we are only interested in an answer to a specific query. This kind of a situation does not require identification of the full function over the whole domain and thus it is not necessary to build the model in the entire area. Approach presented by global methods such as neural networks, neuro-fuzzy networks, polynomial approximation and other in some situation can be successfully replaced by a local modeling methods.

* Corresponding author. Tel.: +4-891-449-5655 ; fax: +4-891-449-5990.
E-mail address: mpietrzykowski@wi.zut.edu.pl

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of KES International
10.1016/j.procs.2017.08.210

The MM-method composes of two groups of algorithms: algorithms for defining the mini-model domain and algorithms for mathematical modeling on the mini-model area. In the previous research the mini-models were developed in several variants. In the base variant MM-domain was described as multidimensional polytope. In the general case for *n*-dimensional space the mini-model domain had a shape of a *n-1*-dimensional convex polytope. The domain was based on: simplex, *n*-cube and *n*-orthoplex. The geometric interpretation is as follows: it is a line segment in a 2D-space, polygon in 3D-space (triangle, quadrilateral, etc.), polyhedron in 4D-space (simplex, hexahedron, etc.). In some previous research the MM-domain had hyper-ellipsoidal shape[6,7,8]. Generally the method can use any approximation algorithm, but in practice simplest method are preferred over the more complex. In most cases linear regression algorithm is appropriate. It is not complicated enough for a mathematical modeling in entire domain of real world dataset. However, we have to remember that this method is used only in some part of the space. In the task of local modeling it is suitable for finding local data dependency. Moreover it provides the ability to local dimensionality reduction for a particular query point[4].

The main aim of the paper is to analyze MM-method based on centroid clustering methods. The author will examine three centroid-based clustering method: *k*-means[16,15], *k*-medoids[18,19], *c*-means[20,23]. The article introduce two new versions of the MM-method which are based on *k*-medoids method and the fuzzy clustering methods - *c*-means algorithm. The variant of the method based on *k*-means clustering algorithm were previously described[9]. The paper also presents extensive comparison of the mini-models method based on centroid clustering algorithm with previous versions of the MM-method. In the presented modification, the domain of the model is calculated using the clustering algorithms. Clustering algorithm makes the MM-method simpler and gives lower computational complexity. In the previous version each query point has its own domain and all domains have to be calculated separately. Clustering algorithms divide the input space at once at the beginning of the learning process. Then, every cluster is treated as separate MM-domain. It gives an edge in the situation when we have to calculate answers for several query points. This approach frees us from creating MM-domains for different query points separately.

The MM-method is a supervised instance-based learning algorithm (memory-based learning). Methods in this group instead of performing explicit generalization, compare new problem instances with instances from training, which are stored in the memory[10,11]. In this type of algorithm complexity can grow with the data. On the other side instance-based learning methods have great ability to adapt its model to previously unseen data. It simply could add a new instance to the memory. Examples of instance-based learning algorithm are the *k*-nearest neighbor algorithm[12,30,31], GRNN network[13,32] or RBF network[14,33]. The *k*-NN method can be considered as a special case of a mini-model method[5].

2. Centroid-based clustering algorithms

There are several different clustering methods type: hierarchical, centroid-based, distribution-based, density-based and other[26,19]. The article only describes some of the centroid-based methods which will be used by the mini-models. In centroid-based clustering, clusters are represented by a central vector. It may not necessarily be a member of the data set. The article will focus on three methods: *k*-means[16,15], *k*-medoids[18,19], *c*-means[20,21]. Clustering algorithms are used for input space division into several subspaces. It is an iterative, data-partitioning procedure that assigns n observations to exactly one of k clusters. The main drawback of these method is k (c) parameter because it is chosen *a priori* before the algorithm starts.

2.1. k-means

In the basic version of *k*-means algorithm the centroid is computed as a mean of all data points belonging to the cluster. The algorithm proceeds as follows:

1. Choose k initial cluster centers (centroids).
2. Compute point-to-cluster-centroid distances of all observations to each centroid.
3. Assign each observation to the cluster with the closest centroid.
4. Compute the average of the observations in each cluster to obtain k new centroid locations.