Data & Knowledge Engineering 101 (2016) 24–41

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

An instance selection method for large datasets based on Markov Geometric Diffusion

Duílio A.N.S. Silva[a], Leandro C. Souza[b], Gustavo H.M.B. Motta[a,*]

[a] Centro de Informática (CI), Universidade Federal da Paraíba (UFPB), João Pessoa, PB, Brazil
[b] Departamento de Ciências Exatas e Naturais (DCEN), Universidade Federal Rural do Semi-árido (UFERSA), Mossoró, RN, Brazil

ARTICLE INFO

Article history:
Received 12 November 2013
Received in revised form 30 March 2015
Accepted 10 November 2015
Available online 19 November 2015

Keywords:
Data mining
Instance selection
Markov geometric diffusion
Large datasets

ABSTRACT

Given the growing amount of data produced from within different areas of knowledge, data mining methods currently have to face challenging datasets with greater numbers of instances and attributes. However, the processing capacity of data mining algorithms is struggling under this growth. One alternative for tackling the problem is to perform instance selection on the data in order to reduce its size, as a preprocessing step for data mining algorithms.

This study presents e-MGD, a method for instance selection as an extension of the Markov Geometric Diffusion method, which is a linear complexity method used in computer graphics for the simplification of triangular meshes. The original method was extended so that it was capable of reducing datasets commonly found in the field of data mining. For this purpose, two essential points of adjustment were required. Firstly, it was necessary to build a geometric structure from the data and secondly, to adjust the method so that it could deal with types of attributes encountered within these datasets. These adjustments however, did not influence the complexity of the final e-MGD, since it remained linear, which enabled it to be applied to datasets with a greater number of instances and features. One distinct characteristic of the proposed extension is that it focuses on preserving dataset information rather than improving classification accuracy, as in the case of most instance selection methods.

In order to assess the performance of the method, we compared it with a number of classical and contemporary instance selection methods using medium to large datasets, plus a further set of very large datasets. The results demonstrated a good performance in terms of classification accuracy when compared to results from other methods, indicating that the e-MGD is a good alternative for instance selection.

1. Introduction

Computing applications from within several areas of knowledge have generated data with a large number of instances. To extract intrinsic knowledge or patterns for such data, there are solutions based on data mining. However, large datasets still pose a challenge to existing methods due to the large volume of data that needs to be processed. Improving data mining algorithms [40] or reducing datasets are approaches which deal with this kind of problem. The first approach is impracticable for some algorithms [22], while the second has a greater impact on improving the performance of data processing [42].

* Corresponding author.
E-mail addresses: duilioanss@gmail.com (D. Silva), leandro.souza@ufersa.edu.br (L. Souza), gustavo@ci.ufpb.br (G. Motta).

http://dx.doi.org/10.1016/j.datak.2015.11.002
0169-023X/

Data reduction is used to reduce the size of datasets whilst retaining representative data. Instance selection (IS) is one way of performing data reduction by decreasing the number of instances. Despite the efforts of numerous studies to deal with IS, most face the same problem as data mining algorithms: they are not applicable to datasets with large numbers of instances. These present algorithms with high computational complexity, at least $O(n^2)$ [9,22], which in turn creates a scaling problem. This leads to an increase in storage requirements and impacts on classification accuracy [14].

The present study is an extension of the Markov Geometric Diffusion (MGD) method [45], which is based on the Diffusion Maps Theory [3,4,11]. MGD performs a probabilistic reduction on triangular meshes, which are used in the field of computer graphics [7]. In the reduction process, the object features represented on the mesh are preserved. This is achieved by retaining the elements that most contribute to its structure. This method presents computational complexity $O(n)$ and is able to control the scalability problem for meshes with large numbers of primitives.

Our extension, termed extended-MGD (e-MDG), aims at expanding the reduction capability of the MGD method so that it may be applied to types of data usually found in the data mining field. For this purpose, two points of adjustment are required. One is to build a geometric structure for this kind of data, and the other is to adjust the method so that it may deal with instances with distinct types of attributes. In the original work of the MGD, data reduction is explored in triangular meshes, which have a geometric structure in data. Thus, we present an algorithm to organize unstructured data within large numbers of instances and attributes, as a preprocessing step.

Labeled or unlabeled datasets may both be processed by the e-MGD. The approach is more interested in preserving information rather than just focusing on improving classification accuracy, as in the case of most IS methods. One of the functions of the method is able to determine the representativeness of each instance. In order to perform this function, it needs to be informed of similarity or dissimilarity notions, such as the distance functions between two instances.

Our work contributes with a method that preserves the most representative data in the instance selection process. Due to its low computational complexity, which was also achieved in the preprocessing step for structuring the data, problems where scalability is critical or relevant to its feasibility, would also benefit from such a method. Moreover, as we will demonstrate in the experiment section, experiments were conducted following the same strategy as in IS methods, evaluating metrics such as classification accuracy and reduction rate. The results achieved from the tested datasets were satisfactory and comparable to those used in comparison methods.

The remainder of this paper is organized as follows. Section 2 discusses work related to the content presented herein. Section 3 explores the theory and application of the MGD method. Section 4 describes the extended-MGD method. Section 5 presents how the experiments were setup, followed by the results of the experiments in Section 6. Section 7 presents discussions regarding our methodology, the achieved results and some limitations. Section 8 presents our conclusions and suggestions for future improvements.

2. Related work

There are many approaches to IS in the literature, for example algorithms based on the Nearest Neighbors (NN) rule, in which an instance is removed if its class and the class of the k nearest neighbors are equal, presuming that these instances will afterwards be correctly classified [51]. There are also a number of other studies with different approaches, which have mainly attempted to eliminate the drawback of having to store all the available training examples in the model: the Condensed Nearest Neighbor (CNN) algorithm [29], the Instance-Based learning algorithms (IB2) (IB3) [1], Iterative Case Filtering algorithm (ICF) [6] and Decremental Reduction Optimization Procedure (DROP1-5)[52].

A similar idea may be applied to clustering, where an instance close to the cluster centers of similar instances could be selected as a representative instance. Clustering algorithms such as k-means and fuzzy k-means clustering are used for this purpose. The Okun and Priisalu [39] reduction is based on a combination of the fuzzy k-means clustering algorithm [5] and two nonnegative matrix factorizations. It associates a set of instances with each cluster rather than a single representative. In Eschrich et al. [17], cluster centers are generated, which are later considered to be the centroids. The quality of the reduced dataset will depend on how the centroids represent the cluster in the data.

Whelan et al. [48] initially presented a strategy for spatio-temporal data reduction based on the k-means clustering method. The cluster centers are used as representatives for data instances, together with those closest to the centers. Secondly, to comprehend datasets of different densities, Le-Khac et al. [34] used the same combination of density-based clustering (DBSCAN) and the Shared Nearest Neighbor algorithm (SNN) as proposed by Ertöz et al. [16], but mainly focusing on the context of data reduction for large datasets.

It was proved that IS belongs to the class of NP-hard problems [26]. Therefore, local search heuristics and metaheuristics such as simulated annealing, tabu search, and evolutionary algorithms may be practical approaches for applying to IS [8,13,19,44]. For instance, in [13], an agent-based population learning algorithm is proposed, which selects prototypes from clusters. The agents represent the tabu search and local search procedures. They are responsible for finding better solutions, i.e. to determine and select the best instances according to a fitness function, which is carried out by estimating the classification accuracy of the selected instances. Despite being a new alternative for IS, plus the fact that its results demonstrated the effectiveness of the approach in terms of improving classification accuracy, the experiments only covered datasets with a reduced number of instances.