Neurocomputing 150 (2015) 513–528

Contents lists available at ScienceDirect

Neurocomputing

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

Confidence ratio affinity propagation in ensemble selection of neural network classifiers for distributed privacy-preserving data mining

Yiannis Kokkinos *, Konstantinos G. Margaritis

Parallel and Distributed Processing Laboratory, Department of Applied Informatics, University of Macedonia, 156 Egnatia str., PO Box 1591, 54006 Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 9 December 2013
Received in revised form
20 July 2014
Accepted 21 July 2014
Communicated by Chennai Guest Editor
Available online 2 October 2014

Keywords:
Neural networks
Ensemble selection
Distributed computing
Privacy-preserving
Data mining

ABSTRACT

We consider distributed privacy-preserving data mining in large decentralized data locations which can build several neural networks to form an ensemble. The best neural network classifiers are selected via the proposed confidence ratio affinity propagation in an asynchronous distributed and privacy-preserving computing cycle. Existing methods usually need a shared to all classifiers dataset, in order to examine the classification accuracy of each pair of classifiers. This process is neither distributed nor privacy-preserving. On the other hand in the proposed distributed privacy-preserving solution the classifiers validate each other in a local way. The training set of one classifier becomes the validation set of the other and vice versa and only partial sums of confidences for the correctly and the falsely classified examples are collected. By locally defining a confidence ratio between each pair of classifiers the well known affinity propagation algorithm finds the most representative ones. The construction is parallelizable and the cost is $O(LN)$ for L classifiers and N examples. A-priori knowledge for the number of best classifiers is not required since in affinity propagation algorithm this number emerges automatically. Experimental simulations on benchmark datasets and comparisons with other pair-wise diversity based measures and other existing pruning methods are promising.

1. Introduction

Distributed data mining tasks [1–4] refer to the discovery of potentially useful patterns from large physically distributed data banks. Collecting large volumes of data to a single location for centralized data mining is usually unfeasible. The reasons for decentralization lay in the huge communication costs, computation costs, network bandwidth, central storage requirements, main memory demands and privacy preservation. Distributed data mining [1–4] is challenging due to the large number of distributed data sources, the dynamic character of data and the privacy-preserving issues that concern the participants. Highly decentralized data analysis can be programmed as large collections of independent processes in grids and distributed infrastructures [2]. Thus the field of distributed data mining focuses on developing efficient algorithms for mining patterns or information from distributed (usually disjoint) datasets without the need to centralizing them and sometimes without the need to reveal them to others [4].

Existing distributed data mining approaches vary. One versatile approach is to keep the disjoint datasets to their locations and perform, in parallel, local data mining to produce local models [5–7]. According to this advanced distributed data mining scenario the local models are those that can be transmitted to a central site that combines them into an ensemble, or global model. A second approach is sub-sampling a representative subset of data from each local site and accumulating these subsets to a central site in order to form a global subset. If this representative subset is close to the overall data distribution, then centralized data mining algorithms can be straightforwardly carried out on it, although in some cases the sub-sampling on huge datasets could produce very large subsets and the original scalability problem will remain [6]. A third approach is to create a meta-learner [8] from the ensemble, sometimes by combining the first and second approaches. Distributed data mining via several meta-learning methodologies [9–12] split the dataset into different sites, train a classifier on each site and then post-train a non-linear combining or pruning scheme for the ensemble members [13–15], by using their prediction outputs from an independent evaluation set.

A fourth approach belongs to fully decentralized distributed data mining algorithms [1,2]. The participating locations can communicate directly with each other in a pair-wise fashion via message passing. Some operational characteristics desired for these highly decentralized data mining algorithms are [1]

* Corresponding author.
E-mail addresses: gkokkinos@uom.gr (Y. Kokkinos), kmarg@uom.gr (K.G. Margaritis).

http://dx.doi.org/10.1016/j.neucom.2014.07.065
0925-2312/

distributed (data stays on each site), scalable (can handle large numbers of data), communication efficient (if only point-to-point messages), lock-free (without locking mechanism for simultaneously broadcasting), asynchronous (without the need of synchronization points), decentralized (without the need of a server) and privacy-preserving (without revealing local data). Privacy concerns are those that restrict the transferring as well as the sharing of the sensitive data. In this work we suggest a neural network ensemble selection strategy that possesses a number of the aforementioned operational characteristics.

Without the privacy-preserving factor, or the asynchronous factor, things are easy and the literature is teeming of different ensemble selection methods. We discuss many of them in Section 2. Privacy-preserving means that data exchange is hindered or restricted and thus a participant location must not acquire any extra knowledge of the other participant's data. So essentially they are not able to read other's data or produce an independent evaluation set, gathered from sub-sampling all locations, and compare their output estimations on it. Asynchronous means the lack of a synchronization mechanism or a central coordination. Hence, the essence of the proposed solution is based on plain asynchronous point-to-point message passing in a mutual validation cycle. That is to say, the basic operation two participants in the ensemble can do is to exchange messages. Such a classical point-to-point one-directional communication is depicted in Fig. 1. We then exploit the possibility of mutual validation for mapping all the individual neural network classifiers based on their local accuracy, by using only simple asynchronous pair-wise message exchanging, like send a classifier and receive performance.

The proposed solution design for distributed privacy-preserving neural network ensemble selection consists of typical training the neural networks inside each location, message passing in a mutual validation cycle (classifiers are exchanged), creating the pair-wise similarities as locally defined confidence ratios that accumulate sums of confidences (we also compare with other pair-wise diversity based similarities), selecting the best classifiers by the affinity propagation clustering algorithm [19] which uses message passing (we also compare with other pruning methods).

Therefore, the contribution of this paper is the proposed confidence ratio affinity propagation which can cope with the previously mentioned operational characteristics that is asynchronous, lock-free communications, scalable, distributed and privacy-preserving. We simply suggest using as key factors confidence ratios as pair-wise classifier similarities, a classical pair-wise computing cycle to locally compute these similarities between all pairs of classifiers, and the well known affinity propagation algorithm (see Section 3). Preliminary experimental results on the last where reported in an earlier work [20]. The present article presents the detailed framework together with extensive comparisons with other pair-wise diversity based measures and other existing pruning methods. We cover the case of employing affinity propagation for distributed ensemble selection in view of the fact that this state-of-the-art algorithm had not been exploited so far. The proposed confidence ratio is by nature a locally defined measure of similarity between two classifiers and depends only on their two local training datasets. It is produced via the plain point-to-point message passing illustrated in Fig. 1. Hence, intrinsically the computations for the confidence ratios are: (1) distributed locally, (2) asynchronous and lock-free, (3) privacy-preserving since they leaving the local data banks intact without the need to share data from one another, (4) decentralized without collecting any data to a central location, (5) independently parallelizable.

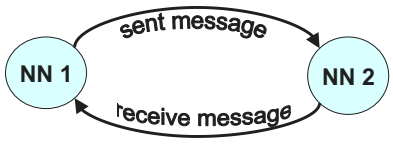

Fig. 1. The classical asynchronous point-to-point one-directional message passing between two locations that hold a different neural network classifier.

In addition, the confidence ratio affinity propagation selects the best k neural network classifiers without the number k to be given in advance, and without the need for monitoring the pruned ensemble performance on a common to all validation set. It is parameter-free and automatically selects the best number of classifiers. Another advantage of the proposed method is the scalability that came from the independently parallel construction of the mutual validation matrix which is a result of the message passing computations. Given L classifiers and N training examples, distributed across L locations where each one holds N/L examples, the computational complexity of constructing the mutual validation matrix is reduced to $O(L^2N/L)=O(LN)$. Therefore the proposed solution is fast and scalable. The results demonstrate that the method automatically manages to select few classifiers and delivers a fast and accurate ensemble without the necessity for additional user input.

The rest of the paper is organized as follows. Section 2 provides short literature review and background knowledge on the basic concepts of distributed privacy-preserving data mining and related work on neural network ensemble selection methods. Section 3 presents in details the proposed ensemble of regularization neural networks, elucidates our procedure of computations for the mutual validation, and describes the ensemble selection via the proposed confidence ratio affinity propagation and the combining via majority voting. Section 4 describes implementation details of many existing pair-wise diversity based measures that we also use as similarities to compare with. Section 5 presents the existing pruning methods we use in the comparisons. Section 6 provides experimental results and comparisons on the effectiveness of the proposed method. Section 7 gives conclusions and summaries.

2. Background material

2.1. Distributed privacy-preserving data mining

A practical definition for 'distributed' is to learn without moving the local data to other locations and for 'privacy-preservation' is to learn without exposing the local data to any other. In a large scale distributed system that composed of several disjoint data banks local data exchange is usually impractical. In addition, the free flow of information is often prohibited by legal obligations or personal concerns. The participants may wish to collaborate but might not fully trust each other. Then distributed privacy-preserving data mining is the how to build valid data mining models and find meaningful patterns without disclosing any private information among the participants. Distributed privacy-preserving data mining is devoted to many real life applications and practical implementations. Few examples of applications [3] are privacy-preservation in personalized systems (newspapers, catalogs, etc.), privacy-preserving medical data mining, genomic privacy, privacy-preserving recommendation systems as well as applications in security-control, intrusion detection and surveillance.

The basic problem to be solved has a simple definition. In a large cooperative environment each participating node has a private input $\mathbf{x}_i$. All nodes wish to collaborative in order to jointly compute the output $f(\mathbf{x}_1, \mathbf{x}_2,\ldots,\mathbf{x}_n)$ of some function f while at the end of the process nothing but the output should have been exposed. Then distributed privacy-preserving data mining solves this problem by allowing nodes to safely share data or extracting