Neurocomputing 147 (2015) 456–471

Contents lists available at ScienceDirect

Neurocomputing

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

Constrained Semi-Supervised Growing Self-Organizing Map

Amin Allahyar [a,*], Hadi Sadoghi Yazdi [a,b], Ahad Harati [a]

[a] Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran
[b] Center of Excellence on Soft Computing and Intelligent Information Processing, Ferdowsi University of Mashhad, Iran

ARTICLE INFO

Article history:
Received 2 July 2012
Received in revised form
30 May 2014
Accepted 21 June 2014
Communicated by B. Hammer
Available online 1 July 2014

Keywords:
Constrained clustering
Online learning
Semi-supervised Self-Organizing Map
Bregman's projection
Metric learning

ABSTRACT

Semi-supervised clustering tries to surpass the limits of unsupervised clustering using extra information contained in occasional labeled data points. However, providing such labeled samples is not always possible or easy in real world applications. A weaker, yet still very useful option is providing constraints on the unlabeled training samples, which is the focus of the Constrained Semi-Supervised (CSS) clustering. On the other hand, online learning has gained considerable amount of interests in real world problems with massive sample size or streaming behavior, as lack of memory and computational resources seriously restrict the application of the offline and batch methods. However, the existing algorithms for online CSS clustering problem either assumed that the entire dataset is available and added constraints incrementally or considered chunks of constrained data points and applied an offline CSS clustering algorithm. Thus, none of them can be categorized as a genuine online CSS clustering algorithm.

In this paper, we propose CS2GS, an online CSS clustering algorithm. CS2GS is constructed by modifying the online learning process of Semi-Supervised Growing Self-Organizing Map, and converting it to an iterative constrained metric learning problem that can be solved using the Bregman's iterative projections. The proposed CS2GS is studied via a series of thorough tests using synthetic and real data including selections from UCI datasets and FEP – a recent bilingual corpus used for sentence aligning stage of machine translation. Experimental results show the effectiveness of CS2GS in online CSS clustering, and prove that indeed, the limits of the system accuracy may be pushed higher using unlabeled samples.

1. Introduction

Clustering and classification are two main fields of research in machine learning. In clustering algorithms, the main goal is to divide the given data samples in groups such that data points in the same group are similar while considerably different from data in other groups. On the other hand, classification algorithms are used when labels are available for training data.[1] A classifier tries to build a predictor so that it can predict correct labels for future unseen data samples. While acquiring new data points in real world problems may be as easy as repeating some measurements, assigning a label to each gathered data would be generally hard [1]. Hence, performing classification is usually costly in realistic scenarios. Alternatively, applying a clustering algorithm is not appropriate because in this case the available knowledge (labels) are discarded completely. Therefore, there is a general trade-off between the cost and performance in such learning systems.

Many papers indicated that even a small amount of extra information about the domain can significantly improve the accuracy of clustering algorithms [1–3]. That is, if a clustering algorithm is capable of effectively utilizing this information in the clustering process, the result will improve accordingly. As the amount of given information increases (which will also increases the computational cost), the algorithm gently shifts its behavior from clustering to classification, resulting in an increase to its accuracy. The main goal in this methodology is to maximize the usage of precious domain knowledge (even if it is incomplete) while keeping the cost of learning as low as possible. This family of learning is called *Semi-Supervised* (SS) clustering which is in the focus of this paper. Another neighbor concept in semi-supervised learning is semi-supervised classification [5]. In this methodology, the main algorithm is a classification approach, but the learning process employs unlabeled data points to improve the result. A schema of these two semi-supervised algorithms is shown in Fig. 1.

In real world problems, user is not always capable of determining the exact label of data points. To overcome this difficulty, the user is asked to provide any extra information about the data points [1]. Generally, such extra information can be given in many

* Corresponding author.
E-mail addresses: Amin.Allahyar@stu-mail.um.ac.ir (A. Allahyar), h-sadoghi@um.ac.ir (H. Sadoghi Yazdi), a.harati@um.ac.ir (A. Harati).

[1] This is true in the traditional classification algorithms. The modern classification algorithms can utilize unlabeled samples along with labeled samples. These algorithms are called semi-supervised classification. More discussion will be given later in this paper.

http://dx.doi.org/10.1016/j.neucom.2014.06.039
0925-2312/© 2014 Elsevier B.V. All rights reserved.

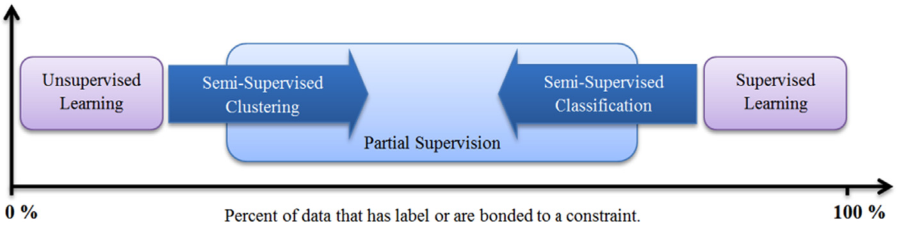
Unsupervised Learning
Semi-Supervised Clustering
Partial Supervision
Semi-Supervised Classification
Supervised Learning
0 %
Percent of data that has label or are bonded to a constraint.
100 %

Fig. 1. Schematic illustration of semi-supervised learning and its two fundamental families of algorithms that are formed by shifting from supervised or unsupervised learning toward the other learning paradigm.

forms. Three most common types are labeled data [6], relative associations [7] and constrained relations [8]. In this paper, the focus is on the constrained relations. They consist of two types: *Must-link* and *Cannot-link*. A must-link constraint is used when two data points should be in a same group while cannot-link constraint is assigned when two data points are required to be in the different groups. Hereafter, *Labeled Semi-Supervised* (LSS) denotes the case of semi-supervised learning with actual labels and semi-supervised learning with the mentioned constraints is called *Constrained Semi-Supervised* (CSS).

Since the last decade, many authors proposed algorithms for CSS learning problem [8–10]. Nonetheless, most of these techniques are offline methods, which assume that a complete set of data points along with their corresponding constraints, are given in advance. However, in real world applications, we often confront a situation where such a complete set is not available or at most, it can be gradually acquired over time. Few existing approaches to the online CSS problem either assumed the availability of the whole dataset with incremental addition of the constraints [11] or considered chunks of data points along with their corresponding constraints and solved the CSS problem with previously proposed offline algorithms [12,13].

In this paper, an online CSS algorithm called Constrained Semi-Supervised GSOM (CS2GS) is proposed. It is based on modification of an existing LSS algorithm called Semi-Supervised Growing Self-Organizing Map (SSGSOM) proposed by Hsu and Halgamuge [6]. CS2GS is capable of incrementally updating its model upon arrival of each pair of data points along with their corresponding constraint that may be available. More specifically, the weight learning problem in SSGSOM is converted to constrained metric learning problem which can be solved using Bregman's iterative projection [14], a popular learning method in online metric learning problems [15–17]. Using this method, we could also prove a limited regret bound on our online CSS learning procedure. We acknowledge that the approach to solve the cost function is only an extension of work done by Davis [15], Kulis [16] and Jain [18]. However, it is the first use of Bregman's iterative projection in the weight learning of a neural network. Furthermore, we suggest improvements in the first layer of SSGSOM algorithm.

The rest of this paper is organized as follows: in Section 2, we will discuss the related work on the online CSS learning. Section 3, gives the preliminaries for the proposed algorithm including the used notations and a brief overview of the SSGSOM algorithm. Section 4 is dedicated to the proposed algorithm. Experimental results are provided in Section 5 for the synthetic data along with some of UCI datasets in addition to FEP, a real-world sentence-aligning data set. Finally, Section 6 encloses the concluding remarks and future works.

2. Related work

As mentioned previously, few studies are dedicated to the incremental CSS problem. The work of Cohn et al. can be considered as one of the early research in the field [13]. It assumes that the resources for storing and processing the whole dataset are available and the dataset is already clustered into groups. The goal in the Cohn's model is to update these clusters based on constraints that are gradually acquired from the user. In the beginning, the current clusters are offered to the user and his feedbacks are collected in the form of new constraints. Next, an offline constrained Expectation Maximization (EM) technique similar to Basu's algorithm [10] is used to form new clusters. This procedure is repeated until the user is satisfied with the clustering results.

Ruiz et al. introduces the concept of stream constraints [19]. Their model, gradually forgets[2] the effects of older data points. Thus, the corresponding constraints are applied for some time and then become obsolete. Ruiz's conceptual model is capable of working with variety of constraints, including must-link and cannot-link. The core clustering algorithm in this model is Constrained *K*-Means proposed by Wagstaff [20]. Although the model is powerful, it forces the data points to satisfy the given constraints. In addition, this work does not provide any explicit cost function or convergence proof. The proposed CS2GS is similar to the Ruiz model as both algorithms gradually forget the past data. However, CS2GS uses a cost function with an analytic solution. Furthermore, as Ruiz algorithm uses a modified Wagstaff's *K*-Means algorithm, it can only form linearly separable clusters while CS2GS can detect clusters which are not linearly separable because it closely mimics the structure of Radial Basis Functions (RBF).

In 2006, Davidson et al. explored the quality of clusters when different constraint sets are applied to single dataset [21]. As expected, in general the average quality of CSS clustering with different sets of constraints is equal to or greater than the average quality in unconstrained version. However, more importantly they showed that there might exist a constraint set that actually decreases the quality of clusters formed by a CSS clustering algorithm. Davidson et al. also defined two measures to calculate the usefulness of a specific constraint set: *Informativeness* and *Coherence*. Informativeness is the amount of information in the constraint set which cannot be extracted from the unconstrained dataset. Coherence measures the amount of agreement of the constraint set with the distance metric [21]. Their studies showed that if the given constraint set has high degree of informativeness and coherence it might greatly improve the final clustering results. Therefore, an incremental CSS clustering algorithm offers remarkable capability for the user as he can selectively provide his constraints according to current clustering results [11].

Davidson et al. extended the Cohn's methodology and defined a new model in which addition or removal of a single constraint would cause the clusters to be updated instead of constructed from scratch [11]. They showed that the solution to this problem is equivalent to the special case of graph coloring problem and thus is NP-Complete. Therefore, they simplified the problem to the case where each data point can be bounded to maximum *m* cannot-link constraints where *m* indicates the number of available clusters.

[2] In this paper, we use "Forget" term to imply the gradual reduction of effect.