Information Sciences 348 (2016) 337–356

Contents lists available at ScienceDirect

Information Sciences

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

Transfer affinity propagation-based clustering

Wenlong Hang[a,1], Fu-lai Chung[b], Shitong Wang[a,b,*]

[a] School of Digital Media, Jiangnan University, Wuxi, Jiangsu 214122, PR China
[b] Department of Computing, Hong Kong Polytechnic University, Hung Hom, Hong Kong, China

ARTICLE INFO

Article history:
Received 4 November 2014
Revised 21 January 2016
Accepted 4 February 2016
Available online 18 February 2016

Keywords:
Transfer learning
Affinity propagation
Exemplars
Transfer affinity propagation
Insufficient datasets

ABSTRACT

Designing a clustering algorithm in the absence of data is becoming a common challenge because the acquisition of annotated information is often difficult or expensive, particularly in the new fields. Because transferring knowledge from the auxiliary domain has been demonstrated to be useful, it is possible to develop an appropriate clustering algorithm for these scenarios in view of transfer learning, where useful information from relevant source domains can be used to complement the decision process and to identify the appropriate number of clusters and a high quality clustering result. In this paper, a novel transfer affinity propagation-based clustering algorithm known as TAP is presented for the scenarios above. Its distinctive characteristics can modify the update rules for the two message propagations used in affinity propagation (AP). Specifically, the most representative points called "exemplars" and the preferences in the source domain are considered for helping in the construction of the high-quality clustering model for insufficient target data. With the corresponding factor graph, the addition of a new term in the objective function for AP allows TAP to cluster in a AP-like message-passing manner for transfer learning, i.e., TAP can identify the appropriate number of clusters and can extract the knowledge of the source domain to enhance the clustering performance for target data, even when the new data are not sufficient to train a model alone. Extensive experiments verify that the proposed algorithm outperforms the state-of-the-art algorithms on insufficient datasets.

1. Introduction

In many real application scenarios, the exponential growth of information has created a strong demand for innovative methods to analyze and manage information. Unfortunately, for some newly emerging or exiguous domains, collected data are usually very sparse, making standard unsupervised clustering algorithms infeasible. Moreover, the acquisition of these insufficient data is usually extremely expensive and time-consuming and involves manual effort. Under this circumstance, unsupervised clustering methods provide very few guarantees because they do not have a high intra-class relationship for target objects. For instance, Fig. 1 shows a motivating example for two types of Aves animals: the familiar chicken and the rare pigeon. The pictures for two types of chicken are given on the left with different categories, and they can be recognized easily because they are relatively common in daily life. Although a pigeon is also one type of Aves animals, the presence of a pigeon is relatively rare and always unusual. Therefore, it becomes a challenging task to identify the categories for a

* Corresponding author at: School of Digital Media, Jiangnan University, Wuxi, Jiangsu 214122, PR China. Tel.: +86 13182791468; fax: +86 85197072.
E-mail addresses: hwl881018@163.com (W. Hang), wxwangst@aliyun.com (S. Wang).
[1] Tel.: +86 18352513348; fax: +86 85197072.

http://dx.doi.org/10.1016/j.ins.2016.02.009
0020-0255/

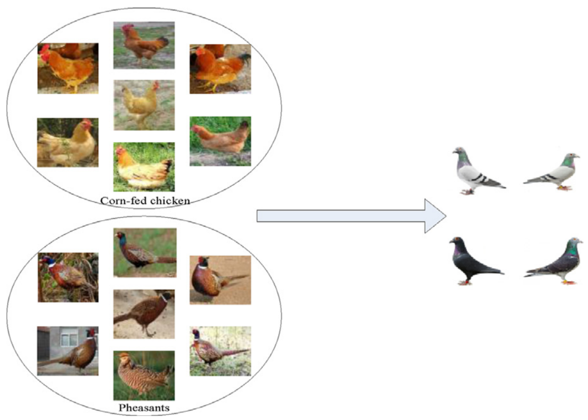

Fig. 1. Two types of Aves animals in which the upper left picture shows the corn-fed chicken, the bottom left picture shows pheasants and the picture on the right shows artificial feeding pigeons and wild pigeons.

pigeon with several pictures if their names are not labeled. This left a couple of questions: (i) Without sufficient data, how many classes are most appropriate? and (ii) Based on prior knowledge, how do we accurately cluster the target objects?

1.1. Related work

To weaken the influence of the insufficient data, two strategies should be taken into consideration: (i) An appropriate clustering algorithm should have the ability to obtain relatively satisfying clustering accuracy; and (ii) A transfer learning based strategy may be used to explore the shared knowledge structure underlying the input domain as the bridge to propagate supervision information from source domain to target domain.

Clustering method: The affinity propagation (AP) [16] clustering algorithm has received considerable attention in the past few years. AP has been widely studied and applied in many fields because it has shown the following distinctive characteristics: 1) Unlike the classical K-Means algorithm or its variants [6], it is not necessary to specify the number of clusters such as the size and structure of the network used in SOM [21]; 2) The most representative points in AP called "exemplars" really exist in the dataset, which is quite different from the other clustering algorithms based on virtual points; 3) AP can obtain the unchangeable clustering results in spite of random initialization; in other words, it is insensitive to the initialization; 4) AP can find more promising clustering results than other exemplar-based clustering algorithms [10,18]. The basic idea of AP is trying to leverage the input similarities between data points to assign them to the most appropriate exemplars. AP uses the max-product belief propagation strategy over factor graph [22] to find the solution of the given objective function, which makes AP an appealing performance. Therefore, it is very attractive and has been developed for different application requirements, such as soft constraint AP [33], hierarchical AP [37], and semi-supervised AP [17].

Despite significant success, one drawback of AP is that it cannot precisely determine the clustering model when target data are insufficient. Because the geometric structure cannot be easily recognized without other auxiliary information in this situation, AP cannot detect the reasonable exemplars. This observation implies that AP is not appropriate for these insufficient datasets, which is also a common technique problem for a large number of clustering algorithms. In reality, it is very common that the amount of recordable data is considerably rare due to the confidentiality of the production or the low yield caused by high cost. Under these circumstances, assignments returned by AP are sensitive to the geometric structure of data points, which mainly occur because AP only aims at minimizing the total distance from all data points to the corresponding exemplars and it loses sight of the underlying characteristics of data points. The main limitation of AP is that it does not simultaneously consider both the minimum energy and other characteristics such as the statistical properties and the geometric structure in the absence of data. As a supplement, transfer learning can effectively remedy this shortage of AP.

Transfer learning: As human beings, all new concepts acquired are not in isolation. However, when considering connections to what is already known, the skill of building analogies is one of the cores of human intelligence [3]. For instance, by focusing only on a visual task, we can always accurately apply prior knowledge in the recognition of a fresh object through cognitive ability. As demonstrated in Fig. 1, corn-fed chickens have a lighter color and a shorter tail than pheasants. We can easily transfer this knowledge to generate a representative hypothesis to recognize the pigeons. This process is known as knowledge transfer, which can make learning additional concepts more efficient. This capacity allows us to mine many types of recurrent patterns and to make inductive inferences for a new task, even when using a small amount of data.

Recently, the literature has shown an increasing interest in developing transfer learning [1,25] algorithms that can be an efficient tool for the target domain in the absence of data [13]. As we may well know, transferred knowledge can be grouped into three broad categories: instances, feature representation, and model parameters. First, the instance transfer approach [7,24,28,40] assumes that certain parts of useful source data can be selected and considered together with available target