Neural Networks 63 (2015) 170–184

Contents lists available at ScienceDirect

Neural Networks

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

Active learning for semi-supervised clustering based on locally linear propagation reconstruction

Chin-Chun Chang *, Po-Yi Lin

Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung, 202, Taiwan

ARTICLE INFO

Article history:
Received 1 May 2014
Received in revised form 4 October 2014
Accepted 14 November 2014
Available online 11 December 2014

Keywords:
Active learning
Semi-supervised clustering
Manifold learning
Locally linear embedding

ABSTRACT

The success of semi-supervised clustering relies on the effectiveness of side information. To get effective side information, a new active learner learning pairwise constraints known as must-link and cannot-link constraints is proposed in this paper. Three novel techniques are developed for learning effective pairwise constraints. The first technique is used to identify samples less important to cluster structures. This technique makes use of a kernel version of locally linear embedding for manifold learning. Samples neither important to locally linear propagation reconstructions of other samples nor on flat patches in the learned manifold are regarded as unimportant samples. The second is a novel criterion for query selection. This criterion considers not only the importance of a sample to expanding the space coverage of the learned samples but also the expected number of queries needed to learn the sample. To facilitate semi-supervised clustering, the third technique yields inferred must-links for passing information about flat patches in the learned manifold to semi-supervised clustering algorithms. Experimental results have shown that the learned pairwise constraints can capture the underlying cluster structures and proven the feasibility of the proposed approach.

1. Introduction

Data clustering algorithms are important tools of data analysis (Jain, Murty, & Flynn, 1999; Xu & Wunsch, 2005). It has been shown that the accuracy of data clustering may be improved by making use of a small amount of side information. Such a kind of data clustering is known as semi-supervised clustering (Grira, Crucianu, & Boujemaa, 2004; Jain, 2010). It is known that the success of semi-supervised clustering relies on effective side information (Davidson, Wagstaff, & Basu, 2006). In Grira, Crucianu, and Boujemaa (2008), Hofmann and Buhmann (1997), Huang and Lam (2009), Klein, Kamvar, and Manning (2002), Nogueira, Jorge, and Rezende (2012) and Wolf, Litwak, Dershowitz, Shweka, and Choueka (2011), active learners are embedded in semi-supervised clustering algorithms because active learning provides means of inquiring useful side information. Active learning and semi-supervised clustering may also be two separated stages (Basu, Banerjee, & Mooney, 2004; Greene & Cunningham, 2007; Mallapragada, Jin, & Jain, 2008; Voevodski, Balcan, Röglin, Teng, & Xia, 2012; Vu, Labroche, & Bouchon-Meunier, 2012; Zhao, He, Ma, & Shi, 2012). Such a kind of active learner is general to many semi-supervised clustering algorithms and is the aim of this paper.

Must-links and cannot-links are the side information to learn in this study. These two kinds of pairwise constraint are defined as that two samples with a must-link should be clustered together, whereas two samples having a cannot-link cannot be in the same cluster. It is also assumed that the must-link defines an equivalence relation among samples. Besides, if samples $\mathbf{x}_i$ and $\mathbf{x}_j$ have a must-link, and samples $\mathbf{x}_j$ and $\mathbf{x}_k$ have a cannot-link, then $\mathbf{x}_i$ and $\mathbf{x}_k$ have an implicit cannot-link as well. Thus, must-links are related to the interior structures of clusters, and cannot-links are about the exterior configurations among clusters.

In Hong and Kwong (2009), it turns out that semi-supervised clustering results may be sensitive to the assignment order of samples. In Basu, Banerjee et al. (2004), to ease semi-supervised clustering, the learned pairwise constraints are structured such that the learned samples belonging to the same cluster are connected together through the learned must-links and known as the *cluster skeleton*. In Greene and Cunningham (2007) and Mallapragada et al. (2008), the learned pairwise constraints are organized in similar manners. However, the three aforementioned active learners have two weaknesses. First, the learned pairwise constraints may bias against some sample subspaces. This is due to selecting queries without considering their importance to the exploration of the sample space. Second, since selecting queries based on

* Corresponding author.
E-mail address: cvml@mail.ntou.edu.tw (C.-C. Chang).

http://dx.doi.org/10.1016/j.neunet.2014.11.006
0893-6080/

random-sampling strategies or the sampling strategies always preferring samples far from learned samples, they may learn little about the interior of clusters when a large number of clusters are present. This fact can be seen by considering that if there are c clusters and every cluster has an equal number n_0 of samples, the probability of two randomly chosen samples belonging to different clusters is $n_0^2 \binom{c}{2} / \binom{cn_0}{2}$, which is equal to $\frac{c-1}{c-1/n_0}$ and at least $\frac{c-1}{c} \underset{c\to\infty}{\longrightarrow} 1$. In addition, samples in different clusters are likely to have long distances between them. Hence, the learned pairwise constraints may be seriously slanted towards cannot-links when c is large.

The proposed approach is also based on the active learning scheme (Basu, Banerjee et al., 2004) for learning structured pairwise constraints. Our approach also exploits the three assumptions for semi-supervised learning, namely, the cluster assumption, the smoothness assumption, and the manifold assumption (Chapelle, Schölkopf, & Zien, 2006). Our active learner has three steps: sample assessment, then cluster exploration, and finally inferred must-link generation.

(1) *Sample assessment*: This step assigns lower query priority to samples unlikely to be in the cluster skeleton. To this end, a kernel version of locally linear embedding (LLE) (Roweis & Saul, 2000) is applied for manifold learning. A new technique called locally linear propagation reconstruction (LLPR) is proposed to determine the importance of a sample to the kernel LLPR of others. The sample neither on a flat patch in the learned manifold nor important to the kernel LLPR of others gets lower priority.
(2) *Cluster exploration*: This step explores the clusters by inquiring the relationship between informative query samples and the learned cluster skeletons. The informative query sample is selected by considering its contribution to the space coverage of the learned cluster skeletons, and the estimated number of queries needed to identify the cluster skeleton of the query sample.
(3) *Inferred must-link generation*: This step expands the learned cluster skeletons through inferred must-links to facilitate semi-supervised clustering. Inferred must-links are yielded based on the three assumptions of semi-supervised learning, and constituted by the samples on flat patches in the learned manifold.

The contributions of this paper are as follows.

- A novel technique based on kernel LLE and kernel LLPR is proposed to determine the query priority of samples. Samples important to local manifold structures are assigned higher priority.
- A novel criterion for query sample selection is proposed. This criterion is in terms of the importance of a sample to expand the space coverage of the cluster skeletons and the number of queries needed to identify the cluster skeleton of the sample.
- Inferred must-links are proposed. Through inferred must-links, information about some flat patches in the learned manifold can be passed to semi-supervised clustering algorithms.
- Two strategies are proposed to attenuate the deficiency of the active learners (Basu, Banerjee et al., 2004; Greene & Cunningham, 2007; Mallapragada et al., 2008) in the case that there exist many clusters. One is considering the number of queries needed to identify the cluster skeleton of a sample in query sample selection. The other is querying samples in regions where samples are difficult to cluster first.

The remaining part of this paper is organized as follows. Section 2 introduces related work. Section 3 presents the proposed LLPR for identifying samples unimportant to cluster structures. Section 4 presents the proposed active learner. Section 5 presents a method of assessing samples for the case that there are many clusters. Section 6 presents the experimental result. Concluding remarks are drawn in the last section.

2. Related work

Side information for semi-supervised clustering may be used for instance-level and space-level implications as Chang and Chen (2012), Grira et al. (2004), Ruiz, Spiliopoulou, and Menasalvas (2010) and references therein show. Instance-level implications may be used to initialize cluster centers, and to guide the clustering process. Space-level implications are applied to induce better distance metrics. Learned pairwise constraints should be useful for both kinds of implication.

In Basu, Banerjee et al. (2004), pairwise constraints are learned by two steps. The first step, known as the exploration step, learns a cluster skeleton for every cluster based on a farthest-first traversal scheme. The second step, known as the consolidation step, consolidates the learned cluster skeletons by learning randomly selected samples. In Mallapragada et al. (2008), the query sample is selected by a min–max criterion, and this algorithm is similar to the exploration step in Basu, Banerjee et al. (2004).

In Greene and Cunningham (2007), ensemble clustering is used to obtain the co-association between samples. Then, cluster skeletons are learned from the representatives of the samples having high co-association. After that, the most uncertain samples are learned to expand the learned cluster skeletons.

In Zhao et al. (2012), based on the concept of the DBSCAN algorithm, samples in the high-density region and samples in the low-density region are referred to as the core samples and the boundary samples, respectively. Then, queries about a core sample and the learned cluster skeletons are followed by two queries regarding the core sample and the nearest and farthest boundary samples of the core sample.

In Vu et al. (2012), a utility measure, known as the ASC score, is defined to measure the potential for two samples being in different clusters. Samples sharing many nearest neighbors are linked to form connected components. Then, pairs of samples in different connected components with no cannot-links are inquired in descending order of the ASC score. In Motta, de Andrade Lopes, and de Oliveira (2009), network centrality measures have also been used to define query order of samples.

In Voevodski et al. (2012), by inquiring one-versus-all queries, accurate clustering results are produced for data sets with unknown metrics. Since not knowing metrics, this work is out of the scope of this study.

The previous studies give us two important insights into active learning for semi-supervised clustering: (1) learning well structured pairwise constraints (Basu, Banerjee et al., 2004; Greene & Cunningham, 2007; Mallapragada et al., 2008) to facilitate semi-supervised clustering; and (2) treating the representative and the uncertain sample differently to ensure that cluster skeletons include representative samples (Greene & Cunningham, 2007; Zhao et al., 2012). However, since gauging the informativeness of a sample only in terms of the distance, the co-association, the number of shared k-nearest neighbors, or the k-neighbor graph, the aforementioned approaches define the importance of samples without considering the space configuration of samples. By using manifold learning, a more specific method is therefore developed in the sequent section.

3. Important sample selection based on locally linear propagation reconstruction

3.1. Locally linear propagation reconstruction

Locally linear embedding (LLE) (Roweis & Saul, 2000) is a popular approach of manifold learning. LLE assumes that data lie on a manifold which can be locally linearly approximated. Denote by $\mathcal{N}_k(\mathbf{x}_i)$ the k-neighbor set of $\mathbf{x}_i$, in which for every sample $\mathbf{x}$ in