Neural Networks 57 (2014) 103–111

Contents lists available at ScienceDirect

Neural Networks

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

Semi-supervised information-maximization clustering

Daniele Calandriello[a,*], Gang Niu[b], Masashi Sugiyama[b]

[a] Politecnico di Milano, Milano, Italy
[b] Tokyo Institute of Technology, Tokyo, Japan

ARTICLE INFO

Article history:
Received 29 October 2013
Received in revised form 18 April 2014
Accepted 26 May 2014
Available online 4 June 2014

Keywords:
Clustering
Information maximization
Squared-loss mutual information
Semi-supervised

ABSTRACT

Semi-supervised clustering aims to introduce prior knowledge in the decision process of a clustering algorithm. In this paper, we propose a novel semi-supervised clustering algorithm based on the information-maximization principle. The proposed method is an extension of a previous unsupervised information-maximization clustering algorithm based on squared-loss mutual information to effectively incorporate must-links and cannot-links. The proposed method is computationally efficient because the clustering solution can be obtained analytically via eigendecomposition. Furthermore, the proposed method allows systematic optimization of tuning parameters such as the kernel width, given the degree of belief in the must-links and cannot-links. The usefulness of the proposed method is demonstrated through experiments.

1. Introduction

The objective of clustering is to classify unlabeled data into disjoint groups based on their similarity, and clustering has been extensively studied in statistics and machine learning. *K-means* (MacQueen, 1967) is a classic algorithm that clusters data so that the sum of within-cluster scatters is minimized. However, its usefulness is rather limited in practice because k-means only produces linearly separated clusters. *Kernel k-means* (Girolami, 2002) overcomes this limitation by performing k-means in a feature space induced by a reproducing kernel function (Schölkopf & Smola, 2002). *Spectral clustering* (Ng, Jordan, & Weiss, 2002; Shi & Malik, 2000) first unfolds non-linear data manifolds based on sample–sample similarity by a spectral embedding method, and then performs k-means in the embedded space.

These non-linear clustering techniques are capable of handling highly complex real-world data. However, they lack objective model selection strategies, i.e., tuning parameters included in kernel functions or similarity measures need to be manually determined in an unsupervised manner. *Information-maximization clustering* can address the issue of model selection (Agakov & Barber, 2006; Gomes, Krause, & Perona, 2010; Sugiyama, Niu, Yamada, Kimura, & Hachiya, 2014), which learns a probabilistic classifier so that some information measure between feature vectors and cluster assignments is maximized in an unsupervised manner. In the information-maximization approach, tuning parameters included in kernel functions or similarity measures can be systematically determined based on the information-maximization principle. Among the information-maximization clustering methods, the algorithm based on *squared-loss mutual information* (SMI) was demonstrated to be promising (Sugiyama, 2013; Sugiyama et al., 2014), because it gives the clustering solution analytically via eigendecomposition.

In practical situations, additional side information regarding clustering solutions is often provided, typically in the form of *must-links* and *cannot-links*: A set of sample pairs which should belong to the same cluster and a set of sample pairs which should belong to different clusters, respectively. Such semi-supervised clustering (which is also known as clustering with side information) has been shown to be useful in practice (Goldberg, 2007; Wagstaff & Cardie, 2000; Wagstaff, Cardie, Rogers, & Schrödl, 2001). *Spectral learning* (Kamvar, Klein, & Manning, 2003) is a semi-supervised extension of spectral clustering that enhances the similarity with side information so that sample pairs tied with must-links have higher similarity and sample pairs tied with cannot-links have lower similarity. On the other hand, *constrained spectral clustering* (Wang & Davidson, 2010) incorporates the must-links and cannot-links as constraints in the optimization problem.

However, in the same way as unsupervised clustering, the above semi-supervised clustering methods suffer from lack of objective model selection strategies and thus tuning parameters included in similarity measures need to be determined manually.

* Corresponding author.
E-mail addresses: daniele.calandriello@mail.polimi.it (D. Calandriello), gang@sg.cs.titech.ac.jp (G. Niu), sugi@cs.titech.ac.jp (M. Sugiyama).

http://dx.doi.org/10.1016/j.neunet.2014.05.016
0893-6080/© 2014 Elsevier Ltd. All rights reserved.

In this paper, we extend the unsupervised SMI-based clustering method to the semi-supervised clustering scenario. The proposed method, called *semi-supervised SMI-based clustering* (3SMIC), gives the clustering solution analytically via eigendecomposition with a systematic model selection strategy. Through experiments on real-world datasets, we demonstrate the usefulness of the proposed 3SMIC algorithm.

2. Information-maximization clustering with squared-loss mutual information

In this section, we formulate the problem of information-maximization clustering and review an existing unsupervised clustering method based on squared-loss mutual information.

2.1. Information-maximization clustering

The goal of unsupervised clustering is to assign class labels to data instances so that similar instances share the same label and dissimilar instances have different labels. Let $\{\boldsymbol{x}_i|\boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^n$ be feature vectors of data instances, which are drawn independently from a probability distribution with density $p^*(\boldsymbol{x})$. Let $\{y_i|y_i \in \{1, \ldots, c\}\}_{i=1}^n$ be class labels that we want to obtain, where c denotes the number of classes and we assume c to be known through the paper.

The information-maximization approach tries to learn the class-posterior probability $p^*(y|\boldsymbol{x})$ in an unsupervised manner so that some "information" measure between feature $\boldsymbol{x}$ and label y is maximized. *Mutual information* (MI) (Shannon, 1948) is a typical information measure for this purpose (Agakov & Barber, 2006; Gomes et al., 2010):

$$\mathrm{MI} := \int \sum_{y=1}^{c} p^*(\boldsymbol{x}, y) \log \frac{p^*(\boldsymbol{x}, y)}{p^*(\boldsymbol{x})p^*(y)} \mathrm{d}\boldsymbol{x}. \tag{1}$$

An advantage of the information-maximization formulation is that tuning parameters included in clustering algorithms such as the Gaussian width and the regularization parameter can be objectively optimized based on the same information-maximization principle. However, MI is known to be sensitive to outliers (Basu, Harris, Hjort, & Jones, 1998), due to the log function that is strongly non-linear. Furthermore, unsupervised learning of class-posterior probability $p^*(y|\boldsymbol{x})$ under MI is highly non-convex and finding a good local optimum is not straightforward in practice (Gomes et al., 2010).

To cope with this problem, an alternative information measure called *squared-loss MI* (SMI) has been introduced (Sugiyama, 2013; Suzuki, Sugiyama, Kanamori, & Sese, 2009):

$$\mathrm{SMI} := \frac{1}{2} \int \sum_{y=1}^{c} p^*(\boldsymbol{x})p^*(y) \left(\frac{p^*(\boldsymbol{x}, y)}{p^*(\boldsymbol{x})p^*(y)} - 1 \right)^2 \mathrm{d}\boldsymbol{x}. \tag{2}$$

Ordinary MI is the *Kullback–Leibler (KL) divergence* (Kullback & Leibler, 1951) from $p^*(\boldsymbol{x}, y)$ to $p^*(\boldsymbol{x})p^*(y)$, while SMI is the *Pearson (PE) divergence* (Pearson, 1900). Both KL and PE divergences belong to the class of the *Ali–Silvey–Csiszár divergences* (Ali & Silvey, 1966; Csiszár, 1967), which is also known as the *f-divergences*. Thus, MI and SMI share many common properties, for example, they are non-negative and equal to zero if and only if feature vector $\boldsymbol{x}$ and label y are statistically independent. Information-maximization clustering based on SMI was shown to be computationally advantageous (Sugiyama et al., 2014). Below, we review the SMI-based clustering (SMIC) algorithm.

2.2. SMI-based clustering

In unsupervised clustering, it is not straightforward to approximate SMI (2) because labeled samples are not available. To cope with this problem, let us expand the squared term in Eq. (2). Then SMI can be expressed as

$$\begin{aligned}\mathrm{SMI} &= \frac{1}{2} \int \sum_{y=1}^{c} p^*(\boldsymbol{x})p^*(y) \left(\frac{p^*(\boldsymbol{x}, y)}{p^*(\boldsymbol{x})p^*(y)} \right)^2 \mathrm{d}\boldsymbol{x} \\ &\quad - \int \sum_{y=1}^{c} p^*(\boldsymbol{x})p^*(y) \frac{p^*(\boldsymbol{x}, y)}{p^*(\boldsymbol{x})p^*(y)} \mathrm{d}\boldsymbol{x} + \frac{1}{2} \\ &= \frac{1}{2} \int \sum_{y=1}^{c} p^*(y|\boldsymbol{x})p^*(\boldsymbol{x}) \frac{p^*(y|\boldsymbol{x})}{p^*(y)} \mathrm{d}\boldsymbol{x} - \frac{1}{2}.\end{aligned} \tag{3}$$

Suppose that the class-prior probability $p^*(y)$ is uniform, i.e.,

$$p(y) = \frac{1}{c} \quad \text{for } y = 1, \ldots, c.$$

Then we can express Eq. (3) as

$$\frac{c}{2} \int \sum_{y=1}^{c} p^*(y|\boldsymbol{x})p^*(\boldsymbol{x})p^*(y|\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \frac{1}{2}. \tag{4}$$

Let us approximate the class-posterior probability $p^*(y|\boldsymbol{x})$ by the following kernel model:

$$p(y|\boldsymbol{x}; \boldsymbol{\alpha}) := \sum_{i=1}^{n} \alpha_{y,i} K(\boldsymbol{x}, \boldsymbol{x}_i), \tag{5}$$

where $\boldsymbol{\alpha} = (\alpha_{1,1}, \ldots, \alpha_{c,n})^\top \in \mathbb{R}^{cn}$ is the parameter vector, $^\top$ denotes the transpose, and $K(\boldsymbol{x}, \boldsymbol{x}')$ denotes a kernel function. Let $\boldsymbol{K}$ be the kernel matrix whose (i, j) element is given by $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and let $\boldsymbol{\alpha}_y = (\alpha_{y,1}, \ldots, \alpha_{y,n})^\top \in \mathbb{R}^n$. Approximating the expectation over $p^*(\boldsymbol{x})$ in Eq. (4) with the empirical average of samples $\{\boldsymbol{x}_i\}_{i=1}^n$ and replacing the class-posterior probability $p^*(y|\boldsymbol{x})$ with the kernel model $p(y|\boldsymbol{x}; \boldsymbol{\alpha})$, we have the following SMI approximator:

$$\widehat{\mathrm{SMI}} := \frac{c}{2n} \sum_{y=1}^{c} \boldsymbol{\alpha}_y^\top \boldsymbol{K}^2 \boldsymbol{\alpha}_y - \frac{1}{2}. \tag{6}$$

Under orthonormality of $\{\boldsymbol{\alpha}_y\}_{y=1}^c$, a global maximizer is given by the normalized eigenvectors $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_c$ associated with the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$ of $\boldsymbol{K}$. Because the sign of eigenvector $\boldsymbol{\phi}_y$ is arbitrary, we set the sign as

$$\widetilde{\boldsymbol{\phi}}_y = \boldsymbol{\phi}_y \times \mathrm{sign}(\boldsymbol{\phi}_y^\top \mathbf{1}_n),$$

where $\mathrm{sign}(\cdot)$ denotes the sign of a scalar and $\mathbf{1}_n$ denotes the n-dimensional vector with all ones. On the other hand, since

$$p^*(y) = \int p^*(y|\boldsymbol{x})p^*(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \approx \frac{1}{n} \sum_{i=1}^{n} p(y|\boldsymbol{x}_i; \boldsymbol{\alpha}) = \boldsymbol{\alpha}_y^\top \boldsymbol{K} \mathbf{1}_n,$$

and the class-prior probability was set to be uniform, we have the following normalization condition:

$$\boldsymbol{\alpha}_y^\top \boldsymbol{K} \mathbf{1}_n = \frac{1}{c}.$$

Furthermore, negative outputs are rounded up to zero to ensure that outputs are non-negative.

Taking these post-processing issues into account, cluster assignment y_i for $\boldsymbol{x}_i$ is determined as the maximizer of the approximation of $p(y|\boldsymbol{x}_i)$:

$$y_i = \operatorname*{argmax}_y \frac{[\max(\mathbf{0}_n, \boldsymbol{K}\widetilde{\boldsymbol{\phi}}_y)]_i}{c\ \max(\mathbf{0}_n, \boldsymbol{K}\widetilde{\boldsymbol{\phi}}_y)^\top \mathbf{1}_n} = \operatorname*{argmax}_y \frac{[\max(\mathbf{0}_n, \widetilde{\boldsymbol{\phi}}_y)]_i}{\max(\mathbf{0}_n, \widetilde{\boldsymbol{\phi}}_y)^\top \mathbf{1}_n},$$