A VARIATIONAL MULTI-VIEW LEARNING FRAMEWORK AND ITS APPLICATION TO IMAGE SEGMENTATION

Zhenglong Li¹, Qingshan Liu¹,² and Hanqing Lu¹

¹National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, P.O. Box 2728, Beijing (100190), P.R. China.
zlli@live.com, {qsliu,luhq}@nlpr.ia.ac.cn
²Depart. Computer Sciences, Rutgers, the State University of New Jersey, 110 Frelinghuysen Road, Piscataway, NJ 08854-8019.
qsliu@cs.rutgers.edu

ABSTRACT

The paper presents a novel multi-view learning framework based on variational inference. We formulate the framework as a graph representation in form of graph factorization: the graph comprises of factor graphs, which are used to describe internal states of views. Each view is modeled with a Gaussian mixture model. The proposed framework has three main advantages: 1) less constraint assumed on data, 2) effective utilization of unlabeled data, and 3) automatic data structure inferring: proper data structure can be inferred in only one round. The experiments on image segmentation demonstrate its effectiveness.

Index Terms— Variational inference, multi-view learning, image segmentation

1. INTRODUCTION

Multi-view learning is a method that utilizes the different description aspects of data. In real world, data can be described from different viewpoints. For instance, we can use different visual cues to represent an image, and one visual cue can be regarded as a ‘view’. In multi-view learning, usually multiple independent hypotheses are learnt in views, in which the unlabeled data are probabilistically labeled by others. Efforts of multi-view learning can be traced back to co-training framework by Blum and Mitchell [1]. In co-training, two classifiers are trained in two different views. The unlabeled data will be labeled by the classifier learnt in one view, and the part of unlabeled data with high confidences will be added to the training data to improve the classifier training in another view. By transferring the most confident labels, two classifiers bootstrap each other in two views. There are three assumptions in co-training: 1) sufficiency: each view is enough to learn a classifier with good performance, 2) compatibility: prediction functions in each view agree on the prediction results of the most data, and 3) independence: views are conditionally independent. Dasgupta et al [2] proved that generalization error of co-training has an upper bound in terms of disagreement between two hypotheses in independent views. Nigam and Ghani [3] proposed an improvement to co-training. Their improvement, i.e. co-EM, can be taken as a multi-view version of Expectation Maximization. There are also some multi-view methods that optimize object function composed of loss term and regularization term, such as Sindhwani et al [4] and Krishnapuram et al [5].

In the paper, we propose a novel variational multi-view learning framework that can effectively employ unlabeled data with small amount of labeled data. We formulate the proposed framework as a graph factorization: the graph comprises of factor graphs, which are used to describe internal states of views. To make optimization related to estimation of the posterior tractable, a variational Bayes approach [6] is employed in the framework, in which a technique of factor analyzer [6, 7] is used to approximate the posterior distribution. We model the internal state of views with Gaussian mixture model: each factor graph is incarnated in Gaussian mixture model. In the optimization, we use a component splitting strategy to introduce new component to avoid local minima and sub-fitting problem. We show an application of the framework to image segmentation. We first label small amount of pixels as foreground and background respectively, and then we use the proposed method to learn the foreground from image.

Compared with previous works, the proposed framework has the following advantages: 1) Less constraint assumed on data. Compared with co-training, the framework does not require that each view should be strictly sufficient, and when encountering incompatible prediction the view with high confidence will be dominant. 2) Effective utilization of unlabeled data. 3) Automatic data structure inferring. For inferring underlying data structure, there exist two kinds of approaches. One is using some stochastic sampling technique, such as Dirichlet process based method [8] that randomly introduces
new component. Albeit the stochastic sampling incurs high computation cost. The other is in the manner of trial-and-error [9], in which several data structures are inferred and the one with highest score according to some metric is chosen. Different from these two strategies, our method is with low computation cost, and a proper data structure can be inferred in only one round.

The rest of the paper is organized as following. In section 2, we will present the proposed variational multi-view learning framework. Experimental results are reported in section 3. Section 4 concludes the paper.

2. VARIATIONAL MULTI-VIEW LEARNING

2.1. Problem Formulation and Proposed Framework

In machine learning, the goal is to predict output \( y \in Y \) given input \( x \in X \), i.e. to learn a map \( f : X \rightarrow Y \), where \( Y = \{1, 2, \cdots, K\} \) and \( X \subset \mathbb{R}^d \) are class labels and the input feature set respectively. Tuple \( (x, y) \) is usually assumed to conform to some underlying distribution \( D \), and the posterior \( p(y|x) \) is estimated for prediction. To learn the posterior, some training data are needed. For semi-supervised learning, the training data are composed of two parts: \( \{x_i, y_i\}^I_1 \) and \( \{x_i', y_i'\}^I_1 \) where \( y_i \) are known and \( y_i' \) unknown respectively.

Let us consider multiple views of the data, \( (x^1, \cdots, x^M, y) \), and the following decomposition is assumed to hold,

\[
p(x | y) = p(x^1, \cdots, x^M | y) = \prod_{m=1}^{M} p(x^m | y),
\]

where \( x = \{x^m\}^{M}_{m=1} \). Denote the parameters of the framework as \( \theta \). The joint distribution of observation \( x \), latent variable \( y \), and the parameters \( \theta \) is given as

\[
p(x, y, \theta) = p(y|\pi)p(\pi) \prod_{m=1}^{M} \left( p(x^m | y, \theta^m)p(\theta^m) \right),
\]

where \( p(y|\pi) \) is the prior for the existence of \( y \).

Fig. (1) shows the directed acyclic graph of the proposed model (2). The parts surrounded by dashed rectangles are factor graphs that represent the internal state of views, corresponding to \( p(x^m | y, \theta^m)p(\theta^m) \) in the right hand side of (2). The whole graph is divided into two parts: instance (represented by latent variable \( y \)) and its different observations (different views). The total observation is the product of factor graphs.

Take account of the marginal likelihood and apply Jensen inequality, we can obtain

\[
\ln p(x) = \ln \int p(x, z) dz \\
> \int q(z) \ln \left( \frac{p(x, z)}{q(z)} \right) dz, \tag{3}
\]

where \( \mathcal{L}(q) \) is the lower bound of \( \ln p(x) \), and \( z \) is shorthand for \( \{y, \theta\} \). In order to maximize \( \ln p(x) \), we maximize the lower bound \( \mathcal{L}(q) \) w.r.t \( q(z) \) in a free form, i.e., variational form, instead of maximizing the log marginal likelihood itself. Note \( \mathcal{L} \) is the negative of Kullback-Leibler (K-L) divergence, and \( \mathcal{L} \) is maximized when \( p(x, z) = q(z) \).

Usually the posterior is intractable. To handle this problem, some approximation is required for \( q(z) \). In the paper, a factorization technique [6, 7] is adopted for the joint distribution between the latent variable and the parameters,

\[
q(z) = q(y, \pi, \theta) = q(y)q(\pi, \theta) = \prod q(z_j) \quad \tag{4}
\]

For simplicity, we denote \( q(z_j) \) as \( q_j \), and rewrite \( \mathcal{L} \) as

\[
\mathcal{L}(q_j) = -\int q_j \ln \frac{q_j}{\tilde{p}(x, z_j)} dz_j + \text{const}, \tag{5}
\]

where \( \tilde{p}(x, z_j) \) is defined as in [10],

\[
\tilde{p}(x, z_j) \equiv C \exp \left( \int \ln p(x, z) \prod_{i \neq j} q_i dz_i \right), \tag{6}
\]

and \( C \) is a normalization constant. Maximizing the lower bound \( \mathcal{L} \) (cf. (5)) w.r.t \( q_j \) is equivalent to minimizing KL\( (q_j || \tilde{p}(x, z_j)) \), i.e.,

\[
q_j \propto \tilde{p}(x, z_j). \tag{7}
\]

2.2. Implementation with Mixture Model

In the proposed framework, we use Gaussian mixture model to represent the internal state, and the detailed dependence
between the parameters and the observation is depicted in Fig. (1) (see the dashed rectangle parts). We model each view $x^m$ with a $K$-component Gaussian mixture with mean $\mu_k^m$, covariance $\Sigma_k^m$, and weights $w_k^m$, and the parameters are given the conjugate priors as following
\[ p(\mu^m | \Lambda^m) = N(\mu^m | \mu_0^m, (\Sigma_k^m \Lambda_k^m)^{-1}), \]
\[ p(\Lambda^m) = \mathcal{W}(\Lambda^m | \mathbf{W}_0, \nu_0^m), \]
\[ p(w^m) = \operatorname{Dir}(w^m | \alpha^m), \]
where $\mathcal{N}(\cdot)$, $\mathcal{W}(\cdot)$, Dir(·) and $\{m_k^m, \alpha_k^m, \beta_k^m, \mathbf{W}_m, \nu_0^m\}$ are Gaussian, Wishart, Dirichlet distributions, and hyperparameters, respectively.

Based on (7), we can easily obtain the posterior [6, 10],
\[ r_{ik} \propto \exp \left\{ \sum_{m=1}^{M} (\psi(\alpha_k^m) - \psi(\sum_{m=1}^{K} \alpha_k^m)) + \frac{1}{2} \sum_{i=1}^{D} \psi \left( \frac{v^m + \frac{1}{2} - i}{2} \right) + \frac{1}{2} \ln \left(2^D | \mathbf{W}_k^m| \right) - \frac{D}{2 \beta_k^m} - \frac{1}{2} \lambda_i \left( \mathbf{x}_i - \mathbf{m}_k^m \right)^T \mathbf{W}_k^m \left( \mathbf{x}_i - \mathbf{m}_k^m \right) \right\}, \tag{8} \]

where $r_{ik}$ represents $p(y = k | \mathbf{x}_i, \ldots, \mathbf{x}_i^m)$ and $r_{ik}$ is also known as ‘responsibility’ that accounts for instance belonging to some class. The parameters in (8) are solved by the following iterations,
\[ \alpha_k^m = \alpha_0^m + N_k^m, \quad \mathbf{m}_k^m = \frac{1}{N_k^m} (\bar{\mathbf{m}}_k^m + N_k \bar{\mathbf{x}}_k^m), \tag{9} \]
\[ \beta_k^m = \beta_0^m + N_k, \quad v_i^m = v_0^m + N_k, \]
\[ (W_k^m)^{-1} = \left( \frac{\beta_0^m + N_k}{\beta_k^m + N_k} \right) \left( \bar{\mathbf{m}}_k^m - \mathbf{m}_0^m \right) \left( \bar{\mathbf{m}}_k^m - \mathbf{m}_0^m \right)^T + (W_0^m)^{-1} + N_k \mathbf{S}_k^m, \tag{10} \]
where $N_k, \bar{\mathbf{m}}_k^m, \mathbf{S}_k^m$ is defined as
\[ N_k = \sum_{i=1}^{N} r_{ik}, \quad \bar{\mathbf{m}}_k^m = \frac{1}{N_k} \sum_{i=1}^{N} r_{ik} \mathbf{x}_i^m, \tag{12} \]
\[ \mathbf{S}_k^m = \frac{1}{N_k} \sum_{i=1}^{N} r_{ik} (\mathbf{x}_i^m - \bar{\mathbf{m}}_k^m) (\mathbf{x}_i^m - \bar{\mathbf{m}}_k^m)^T. \tag{13} \]

In the optimization of (8)–(11), it is possible to find $N_k$ approaches 0. For the component $k$ with $N_k = 0$, there are no sufficient observations to support them, and they can be removed without influencing $\mathcal{L}$. Although by such component elimination a proper component number can be assigned to the underlying optimal structure of the data, it is prone to be stuck into a local minima. Inspired by the works of [11, 7], we develop a component splitting strategy to overcome the local minimum, and introduce new components into the current structure: for each component $k$, $\mathcal{L}(q(z_k))$ is computed, and according to $\mathcal{L}(q(z_k))$, the one with the largest contribution to the lower bound is chosen to be split into two components; an alternative is to choose the one randomly with a probability proportional to the normalized $\mathcal{L}(q(z_k))$. In the paper, we use the former to do component splitting.

3. EXPERIMENTS

We applied the framework to image segmentation. In the experiments, we used two views, i.e., two features, for one image:

View 1 is simply $L* a* b^*$ color value of each pixel, i.e., a 3-dimensional vector.

View 2 is composed of two parts: textural descriptor and spatial coordinate. The textural descriptor comprises of local standard variation $\sigma$ and local entropy $\epsilon$. Local standard variation at pixel $i$ in image $I$ is as following
\[ \sigma_i = \frac{1}{9} \sqrt{\sum_{j \in S_i} (I_j - \bar{I}_S)^2}, \]
where $I$ stands for a grayscale image, $i$ is pixel index in image plane, $S_i$ is a $9 \times 9$ neighborhood centered at $i$ and $\bar{I}_S$ is the mean value of $S_i$. And local entropy is given as
\[ e_i = \sum_{j \in S_i} h_{S_i}(I_j) \log(h_{S_i}(I_j) + \delta), \]
where $h_{S_i}$ is the histogram of neighborhood $S_i$, and $\delta$ is a small number to avoid singularity. The remainder of the feature is the 2-dimensional Cartesian coordinate that is for the sake of spatial coherence. The resulting view 2 is 4-dimensional.

In the experiment, we first label a small amount of pixels as foreground and background respectively. The foreground is modeled with 5 Gaussian components, and so does the background. Before utilizing the unlabeled pixels, k-means clustering is applied to the labeled foreground and background to get the initial parameters for the Gaussian mixture models respectively. The utilization of the unlabeled data is straightforward. During the learning process, in each view the unlabeled pixels will be given responsibilities for background and foreground. And only when two views achieve same opinion on the prediction or one view is dominant over the other, the optimization will converge. In practice the framework will converge very quickly. And unlike the co-training there is no requirement that each view should be both sufficient: only one sufficient view is enough.

Fig. 2 shows the segmentation results by the proposed method and the conventional EM. In Fig. 2(a) the pixels labeled by red is the foreground, and blue the background. Those labeled pixels are used as initial training data. The results of the proposed framework are illustrated in Fig. 2(b). We can see that the foregrounds are well learnt from the
image. In Fig. 2.(c), the results by the conventional EM are presented. For the EM only single view 2 is used, and the labeled pixels are used for learning initial estimate of the foreground. It can be found that the conventional EM failed with the setting of small amount of the labeled data and a large number of the unlabeled data. It is obvious that the proposed method can effectively employ the unlabeled data in a transductive manner to accomplish the task. The success of the proposed method is due to two aspects: 1) the different view utilization of the proposed method effectively employs the unlabeled data, and 2) automatic structure inferring can avoid over-fitting or sub-fitting with a proper component determination.

![Fig. 2](image)

Fig. 2. Experimental results. (a) Original image (red for foreground and blue for background. (b) Results by the proposed method. (c) Results by EM.

4. CONCLUSIONS

In the paper we presented a variational multi-view learning framework in which variational inference is used to deduce the model. The proposed framework is formulated as a graph representation in form of graph factorization: the graph comprises of factor graphs that stand for internal states of views. The Gaussian mixture is used to model the internal state, i.e., each view is modeled by a Gaussian mixture model. We tested the proposed method for image segmentation, and the experimental results demonstrated the effectiveness of the proposed method.

5. ACKNOWLEDGEMENT

We would like to acknowledge the supports from Natural Sciences Foundation of China under grant No. 60835002, 60605004 and 60723005.

6. REFERENCES


