A Variational Approach to Semi-Supervised Clustering

Peng Li, Yiming Ying and Colin Campbell

Department of Engineering Mathematics, University of Bristol,
Bristol BS8 1TR, United Kingdom

Abstract. We present a Bayesian variational inference scheme for semi-supervised clustering in which data is supplemented with side information in the form of common labels. There is no mutual exclusion of classes assumption and samples are represented as a combinatorial mixture over multiple clusters. We illustrate performance on six datasets and find a positive comparison against constrained K-means clustering.

1 Introduction

With semi-supervised clustering the aim is to find clusters or meaningful partitions of the data, aided by the use of *side information*. This side information can be in the form of *must-links* (two samples must be in the same cluster) and *cannot-links* (two samples must not belong to the same cluster). A number of approaches have been proposed for semi-supervised clustering including modifications of pre-existing clustering schemes such as incremental clustering algorithms [1] or K-means clustering [2]. One problem with some of these approaches is that the method can work well if the correct number of clusters is already known: K-means clustering is an example. However, a principled approach to finding K is generally not given. Other potential disadvantages of some clustering approaches is that there is an implicit mutual exclusion of clusters assumption. In many applications this assumption may not be fully valid and it would be more appropriate for a sample to be associated with several clusters.

Our motivation for considering semi-supervised clustering comes from potential applications in cancer informatics. There have been a number of instances where unsupervised learning methods have been applied to cancer expression array datasets and clinically distinct cancer subtypes have been indicated e.g. [3, 4]. However, in some cases a specific causative event is known and thus it is possible to give common labels to some samples. In these cancer applications, the side information is typically in the form of *must-links*: these will be the focus of this paper. We thus propose a probabilistic graphical model approach to semi-supervised clustering with samples represented as combinatorial mixtures over a set of soft clusters. By allowing a representation overlapping different clusters we can derive a confidence measure for cluster membership. In clinical cancer applications, a subtype assignment confidence measure is plainly important. In the next section we propose our probabilistic graphical model, followed by experiments in section 3.

2 The Method

Our semi-supervised model utilizes the Latent Process Decomposition (LPD) model developed in [5, 6], and hence we will call this variant semi-supervised LPD or SLPD. For the natural numbers we adopt the notation $\mathbb{N}_m = \{1, \ldots, m\}$ for any $m \in \mathbb{N}$. For the data we use d as the sample index, g as the attribute index, and script letters $\mathcal{D}, \mathcal{G}$ to index the corresponding number of samples and attributes. The number of clusters is $\mathcal{K}$. The complete data set is $E = \{E_{dg} : d \in \mathbb{N}_{\mathcal{D}}, g \in \mathbb{N}_{\mathcal{G}}\}$. This notation stems from our cancer informatics motivation of expression value of gene g in sample d.

In our semi-supervised setting, we have additional block information $\mathcal{C}$ where each *block* c denotes a set of data points that is known to belong to a single class. In keeping with standard Bayesian models, we also assume both blocks and the data points in each block are *i.i.d.* sampled. Specifically, this side information can be represented by a $\mathcal{D} \times C$ matrix δ with its entities δ_{dc} defined as follows

$$\delta_{dc} = \begin{cases} 1 & \text{if data } d \text{ is a member of block } c, \\ 0 & \text{otherwise.} \end{cases}$$

In probabilistic terms, the dataset E will be partitioned into $\mathcal{K}$ soft clusters described as follows. For a complete data set, a Dirichlet prior distribution for the distribution of clusters is defined by a $\mathcal{K}$-dimensional parameter α. For each known block c, a distribution θ_c over the set of mixture components indexed by k is drawn from a single Dirichlet distribution parameterized by α. Then, for all samples d in block c (i.e. $\delta_{dc} = 1$), the latent indicator variable Z_{dg} indicates which cluster k is chosen, with probability θ_{ck}, from the common block-specific distribution θ_c. The value E_{dg} for attribute g in sample d is then drawn from the kth Gaussian with mean μ_{gk} and deviation σ_{gk}, denoted as $\mathcal{N}(E_{dg}|\mu_{gk}, \sigma_{gk})$. We repeat the above procedure for each block in $\mathcal{C}$. The graphical model is illustrated in Figure 1 which is motivated by Latent Dirichlet Allocation [5].

The *model parameters* are $\Theta = (\mu, \sigma, \alpha)$ and we use the notation $d \sim c$ to denote sample d in block c. From the graphical model in Figure 1, we can formulate the block-specific joint distribution of the observed data E and the latent variables Z by

$$p(E, \theta, Z|\Theta, \mathcal{C}) = \prod_c p(\theta_c|\alpha) \prod_{d \sim c} p(E_d, Z|\theta_c, \Theta), \tag{1}$$

where $p(\theta_c|\alpha)$ is Dirichlet defined by $p(\theta_c|\alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_{ck}^{\alpha_k - 1}$. Using the block matrix δ, we further see that

$$\begin{aligned} \prod_{d \sim c} p(E_d, Z|\theta_c, \Theta) &= \prod_d \Big[p(Z_d|\theta_c) p(E_d|Z_d, \mu, \sigma) \Big]^{\delta_{dc}} \\ &= \prod_d \prod_g \Big[p(Z_{dg}|\theta_c) \mathcal{N}(E_{dg}|\mu_g, \sigma_g, Z_{dg}) \Big]^{\delta_{dc}}. \end{aligned} \tag{2}$$

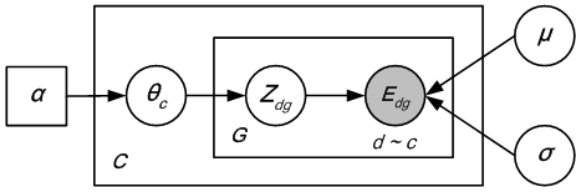

Fig. 1: A graphical model representation of the method proposed in this paper. E_{dg} denotes the value of attribute g in sample d. μ and σ are model parameters. Z_{dg} is a hidden variable giving the cluster index of attribute g in sample d. θ_c gives the mixing over subgroups for sample d in block c denoted by $d \sim c$. The probability of θ_c is given by a Dirichlet distribution with hyper-parameter α.

For notational simplicity, we regard Z_{dg} as a unit-basis vector $(Z_{dg,1}, \ldots, Z_{dg,\mathcal{K}})$ which transforms the cluster latent variable $Z_{dg} = k$ to the unique vector Z_{dg} given by $Z_{dg,k} = 1$ and $Z_{dg,j} = 0$ for $j \neq k$. Equivalently, the random variable Z_{dg} is distributed according to a multinomial probability defined by $p(Z_{dg}|\theta_c) = \prod_k \theta_{ck}^{Z_{dg,k}}$. Hence, the above equation can be rewritten as

$$p(E, \theta, Z|\Theta, \mathcal{C}) = \prod_c p(\theta_c|\alpha) \prod_d \prod_g \prod_k \Big[\theta_{ck} \mathcal{N}(E_{dg}|\mu_{gk}, \sigma_{gk})\Big]^{Z_{dg,k}\delta_{dc}}. \quad (3)$$

With these priors, the final data likelihood can be obtained by marginalizing out the latent variables θ and $Z := \{Z_{dg} : d \in \mathbb{N}_{\mathcal{D}}, g \in \mathbb{N}_{\mathcal{G}}\}$

$$p(E|\Theta, \mathcal{C}) := \int_\theta \sum_Z p(E, \theta, Z|\Theta, \mathcal{C}) d\theta. \quad (4)$$

In particular, we can see from equations (2) and (3) that

$$\begin{aligned} p(E|\Theta, \mathcal{C}) &= \prod_c \int_{\theta_c} \sum_Z \prod_d \prod_g \Big[p(Z_{dg}|\theta_c) \mathcal{N}(E_{dg}|\mu_g, \sigma_g, Z_{dg})\Big]^{\delta_{dc}} p(\theta_c|\alpha) d\theta_c \\ &= \prod_c \int_{\theta_c} \sum_{Z_{dg}, d\sim c} \prod_{d\sim c, g, k} \Big[\theta_{ck} \mathcal{N}(E_{dg}|\mu_{gk}, \sigma_{gk})\Big]^{Z_{dg,k}} p(\theta_c|\alpha) d\theta_c \\ &= \prod_c \int_{\theta_c} \prod_{d,g} \Big[\sum_k \theta_{ck} \mathcal{N}(E_{dg}|\mu_g, \sigma_g, Z_{dg})\Big]^{\delta_{dc}} p(\theta_c|\alpha) d\theta_c. \end{aligned} \quad (5)$$

We should mention that, without block information (i.e. $\delta = I_{\mathcal{D}\times\mathcal{D}}$), the above equation is the exact likelihood of Latent Process Decomposition given in [6].

We now consider model inference and parameter estimation under SLPD. The main inferential goal is to compute the posterior distribution of the hidden variables $p(\theta, Z|E, \Theta, \mathcal{C})$. One direct method is to use Bayes rule $p(\theta, Z|E, \Theta, \mathcal{C}) =$

$\frac{p(E,\theta,Z|\Theta,\mathcal{C})}{p(E|\Theta,\mathcal{C})}$. This approach is usually intractable since this involves computationally intensive estimation of multi-integrals in the final likelihood $p(E|\theta,\mathcal{C})$. In this paper, we rely on *variational inference* methods [7] which maximize a lower bound on the likelihood $p(E|\Theta,\mathcal{C})$ to estimate the model parameters Θ and approximate $p(\theta, Z|E,\Theta,\mathcal{C})$ in a *hypothesis family*. One common hypothesis family is the *factorized family* defined by $q(\theta, Z|\gamma, Q) = q(\theta|\gamma)q(Z|Q)$ with *variational parameters* γ, Q where, in the expression of the distribution q, the dependency on the $E,\Theta,\mathcal{C}$ is omitted. More specifically, in our model we assume that $q(\theta|\gamma) = \prod_c q(\theta_c|\gamma_c) = \prod_c \big(\frac{\Gamma(\sum_k \gamma_{ck})}{\prod_k \Gamma(\gamma_{ck})} \prod_k \theta_{ck}^{\gamma_{ck}-1}\big)$, and $q(Z|Q) = \prod_{d,g} q(Z_{dg}|Q_{dg}) = \prod_{d,g} \big(\prod_k Q_{dg,k}^{Z_{dg,k}}\big)$, among which γ, Q will be set as we describe below. We can lower bound the log-likelihood by applying Jensen's inequality to equation (4):

$$\begin{aligned}\log p(E|\Theta,\mathcal{C}) &= \log \int_\theta \sum_Z p(E,\theta,Z|\Theta,\mathcal{C}) d\theta \\ &\geq \mathcal{L}(\gamma, Q; \Theta) := \mathbb{E}_q\big[\log p(E,\theta,Z|\Theta,\mathcal{C})\big] - \mathbb{E}_q\big[q(\theta,Z|\gamma,Q)\big].\end{aligned}$$

Consequently we can estimate the variational and model parameters by alternative coordinate ascent methods known as a variational EM algorithm:

- E-step: maximize $\mathcal{L}$ with respect to the variational parameters γ, Q to give the following updates ($\Psi(x)$ is the digamma function):

$$Q_{dg,k} = \frac{\mathcal{N}(E_{dg}|\mu_{gk},\sigma_{gk})\big[\prod_c \exp(\delta_{dc}(\Psi(\gamma_{ck}) - \Psi(\sum_k \gamma_{ck})))\big]}{\sum_k \mathcal{N}(E_{dg}|\mu_{gk},\sigma_{gk})\big[\prod_c \exp(\delta_{dc}\Psi(\gamma_{ck}) - \Psi(\sum_k \gamma_{ck}))\big]}, \quad (6)$$

  and

$$\gamma_{ck} = \alpha_k + \sum_{d,g} \delta_{dc} Q_{dg,k} \quad (7)$$

- M-step: maximize $\mathcal{L}$ with respect to μ, σ and α to give:

$$\mu_{gk} = \frac{\sum_d Q_{dg,k} E_{dg}}{\sum_d Q_{dg,k}}, \quad \sigma_{gk}^2 = \frac{\sum_d Q_{dg,k}(E_{dg} - \mu_{gk})^2}{\sum_d Q_{dg,k}}. \quad (8)$$

  with the parameter α found using an additional Newton-Raphson method (see Appendix A in [5] for details).

The above iterative procedure is run until convergence (plateauing of the lower bound on an estimated likelihood $p(E|\Theta,\mathcal{C})$, see [6]). Interpretation of the resultant model is very similar to Latent Process Decomposition [6]. When normalized over k, the parameter γ_{ck} gives the confidence that a sample belonging to block c (which share a common label) belongs to soft cluster k. For each soft cluster k the model parameters μ_{gk} and σ_{gk} give a density distribution for the attribute value g over all samples, see [8] for examples of use of these density estimators in application to interpreting breast cancer array data. If some values

Data	UKM	CKM		ULPD	SLPD	
Sets	0	25%	50%	0	25%	50%
Letter	0.501±0.005	0.502±0.010	0.501±0.009	0.519±0.025	0.521±0.031	**0.527±0.039**
Wine	0.877±0.052	0.885±0.051	0.893±0.047	0.930±0.032	0.926±0.032	**0.935±0.032**
Iris	0.824±0.036	0.825±0.035	0.828±0.041	0.872±0.037	0.910±0.043	**0.920±0.038**
Digit	0.751±0.068	0.758±0.069	**0.772±0.078**	0.736±0.046	0.747±0.045	0.755±0.045

Table 1: A comparison of constrained K-means clustering and SLPD. The entries are BRI (over the 100 trials using 3-fold cross validation). Hypothesis testing indicates a statistically significant performance gain over CKM. The highest BRI score per dataset is given in boldtype.

Data	UKM	CKM		ULPD	SLPD	
Sets	0	25%	50%	0	25%	50%
Leukemia	0.786±0.065	0.782±0.061	0.798±0.062	0.838±0.053	0.846±0.049	**0.851±0.048**
Lung Cancer	0.578±0.030	0.583±0.033	0.599±0.041	0.660±0.033	0.665±0.032	**0.670±0.039**

Table 2: A comparison of constrained K-means clustering and SLPD. The entries are the BRI (mean ± standard deviation over the 50 trials using 3-fold cross validation). Hypothesis testing indicates a statistically significant performance gain over CKM.

of E_{dg} are missing, we omit corresponding contributions in the M-step updates and the corresponding $Q_{dg,k}$. The above argument is based on a maximum likelihood approach. However, following our original argument [6], we can readily formulate a maximum a posterior (MAP) solution penalising over-complex models which fit to noise in the data and we can perform model selection (to find the number of clusters) using hold-out data [8].

3 Experimental Results

To validate the proposed approach, we investigated the ML solution applied to four datasets from the UCI Repository [9] and two cancer expression array datasets. The four data sets from the UCI Repository have known sample labels and thus we can evaluate an objective performance measure. As mentioned, our interest in semi-supervised clustering stems from a potential use in cancer informatics. Thus we consider two datasets for cancer: leukemia array data [3] in which some labels are known due to causative genetic translocations or rearrangements such as the gene fusion events *BCR-ABL* or *E2A-PBX1*. The second dataset is for lung cancer [4]. We investigated three issues. Firstly, a comparison against pre-existing semi-supervised clustering methods, specifically constrained K-means clustering (CKM)[2], since this method is widely used. Secondly, we considered the gains to be made as available label information increases. Finally, we compared unsupervised ULPD with SLPD to evaluate the gains made by using side information. We use the Balanced Rand Index (BRI) as evaluation criterion [10].

In Table 1 we tabulate performance for both constrained K-means clustering and SLPD using BRI with 3-fold random partitioning (one fold being test data, the rest training). Exactly the same sample allocations were used in the evaluation of both constrained K-means clustering and SLPD. For the training set we also imposed a degree of supervision to compare unsupervised learning with semi-supervised clustering using different levels of supervision. The fraction of the supervised data was 0% (unsupervised), 25% and 50%. As observed from Tables 1 and 2, SLPD compares favorably with CKM. With enforcement of a degree of supervision (from 0% to 50%), performance improvement of both SLPD and CKM over LPD and unconstrained K-means clustering (UKM) is observed. As a real-life application we tested SLPD on two cancer expression array datasets for leukemia and lung cancer (Table 2). For leukemia we find that the BRI index improves as we increase the extent of supervision. We also find that SLPD consistently outperforms K-means clustering. A similar picture is repeated for lung cancer.

References

[1] K. Wagstaff and C. Cardie. Clustering with instance-level constraints. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 1103–1110, 2000.

[2] S. Basu, A. Banerjee, and R. J. Mooney. Semi-supervised clustering by seeding. In *Proceedings of International Conference on Machine Learning 2002*, pages 27–34, 2002.

[3] E-J Yeoh et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1:133–143, 2002.

[4] A Bhattacharjee et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences*, 98:13790–13795, 2001.

[5] D. M. Blei, Andrew Y.Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[6] S. Rogers, M. Girolami, C. Campbell, and R. Breitling. The latent process decomposition of cDNA microarray datasets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2:143–156, 2005.

[7] M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

[8] L. Carrivick, S. Rogers, J. Clark, C. Campbell, M. Girolami, and C. Cooper. Identification of prognostic signatures in breast cancer microarray data using bayesian techniques. *Journal of the Royal Society: Interface*, 3:367–381, 2006.

[9] C.L. Blake D.J. Newman, S. Hettich and C.J. Merz. UCI repository of machine learning databases, http://www.ics.uci.edu/∼mlearn/mlrepository.html, 1998.

[10] E. Xing, A. Ng, M. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *Advances in NIPS*, number vol. 15, 2003.