On a Theory of Nonparametric Pairwise Similarity for Clustering: Connecting Clustering to Classification

Yingzhen Yang Feng Liang Shuicheng Yan Zhangyang Wang Thomas S. Huang. UIUC, NUS.

Introduction

Pairwise clustering methods partition the data space into clusters by the pairwise similarity between data points.

The success of pairwise clustering largely depends on the pairwise similarity function defined over the data points. How to design the pairwise similarity in a principled way?

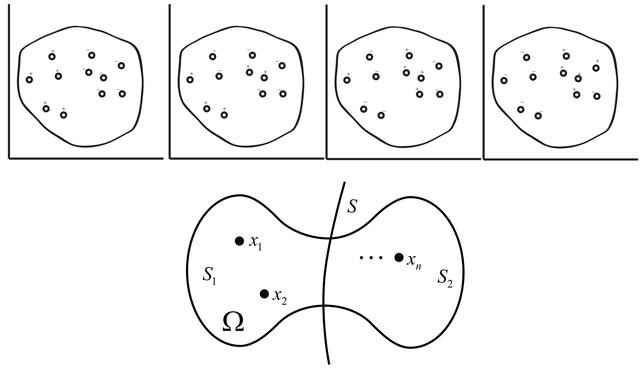
1. Most pairwise clustering methods assume that the pairwise similarity is given (such as kernel similarity), or they learn a more complicated similarity measure based on several given base similarities.
2. **We present a novel pairwise clustering framework by bridging the gap between clustering and multi-class classification.**
3. This pairwise clustering framework learns an unsupervised nonparametric classifier from each data partition, and search for the optimal partition of the data by minimizing the generalization error of the learned classifiers associated with the data partitions. The generalization error bounds are expressed as sum of pairwise similarity.
4. Unlike unsupervised SVM and other information theory based methods that have to estimate parameters of complicated models, the derived pairwise similarity is nonparametric which eliminates the need of parameter estimation.

Contributions

Our main contributions are

1. We derive the generalization error bounds for two unsupervised nonparametric classifiers, i.e. the nearest neighbor classifier and the plug-in classifier, which are the sum of nonparametric pairwise similarity terms between the data points for the purpose of clustering. Under uniform distribution, the nonparametric similarity terms induced by both unsupervised classifiers exhibit a well known form of kernel similarity.
2. We prove that the generalization error bound for the unsupervised plug-in classifier is asymptotically equal to the weighted volume of cluster boundary for Low Density Separation, a widely used criteria for semi-supervised learning and clustering.
3. we propose a new nonparametric exemplar-based clustering method with enhanced discriminative capability, whose superiority is evidenced by the experimental results. The new clustering method uses efficient belief propagation method for label inference.

Formulation



- Given the data $\{\mathbf{x}_l\}_{l=1}^n \overset{i.i.d.}{\sim} P_X, \{\mathbf{x}_l\}_{l=1}^n \subset \mathbb{R}^d$, clustering is equivalent to searching for the optimal hypothetical labeling $\{\mathbf{y}_l\}$.
- For each hypothetical labeling, a non-parametric classifier is learned from the training data $S = \{\mathbf{x}_l, \mathbf{y}_l\}$. In case of the plug-in classifier:

$$\mathrm{PI}_S(X) = \arg\max_{1\le i\le Q} \hat{\eta}^{(i)}_{n,h_n}(X) \quad \hat{\eta}^{(i)}_{n,h_n}(x) = \frac{\sum_{l=1}^n K_{h_n}(x-\mathbf{x}_l)\mathbb{I}_{\{\mathbf{y}_l=i\}}}{n\hat{f}_{n,h_n}(x)}$$

- With high probability, the generalization error of the plug-in classifier satisfies

$$R(\mathrm{PI}_S) \le \hat{R}_n(\mathrm{PI}_S) + \mathcal{O}\Big(\sqrt{\frac{\log h_n^{-1}}{nh_n^d}} + h_n^\gamma\Big)$$

where $\hat{R}_n(\mathrm{PI}_S) = \frac{1}{n^2}\sum_{l,m}\theta_{lm}G_{lm,\sqrt{2}h_n}$, $\theta_{lm} = \mathbb{I}_{\{\mathbf{y}_l\neq\mathbf{y}_m\}}$ is a class indicator function.

$$G_{lm,h} = G_h(\mathbf{x}_l,\mathbf{x}_m),\ G_h(x,y) = \frac{K_h(x-y)}{\hat{f}^{\frac{1}{2}}_{n,h}(x)\hat{f}^{\frac{1}{2}}_{n,h}(y)},$$

- If the nearest neighbor classifier is learned from the hypothetical labeling, with a high probability, its generalization error bound is

$$R(\mathrm{NN}_S) \le \hat{R}_n(\mathrm{NN}_S) + c_0\left(\sqrt{d}\right)^\gamma n^{-d_0\gamma} + \mathcal{O}\Big(\sqrt{\frac{\log h_n^{-1}}{nh_n^d}} + h_n^\gamma\Big)$$

where $\hat{R}_n(\mathrm{NN}) = \frac{1}{n}\sum_{1\le l<m\le n} H_{lm,h_n}\theta_{lm}$,

$$H_{lm,h_n} = K_{h_n}(\mathbf{x}_l-\mathbf{x}_m)\Big(\frac{\int_{\mathcal{V}_l}\hat{f}_{n,h_n}(x)\,\mathrm{d}x}{\hat{f}_{n,h_n}(\mathbf{x}_l)} + \frac{\int_{\mathcal{V}_m}\hat{f}_{n,h_n}(x)\,\mathrm{d}x}{\hat{f}_{n,h_n}(\mathbf{x}_m)}\Big),$$

Formulation Continued

We also show the connection between the derived generalization error bound and Low Density Separation. For the kernel bandwidth sequence $\{h_n\}_{n=1}^\infty$ such that $\lim_{n\to\infty} h_n = 0$ and $h_n > n^{-\frac{1}{4d+4}}$, with probability 1,

$$\lim_{n\to\infty}\frac{\sqrt{\pi}}{2h_n}\hat{R}_n(\mathrm{PI}_S) = \int_S f(s)ds$$

. We can see that the above bound is asymptotically equal to the weighted volume of cluster boundary for Low Density Separation.

Application

- We propose a nonparametric exemplar-based clustering algorithm using the derived nonparametric pairwise similarity by the plug-in classifier. In exemplar-based clustering, each $\mathbf{x}_l$ is associated with a cluster indicator e_l ($l \in \{1,2,...n\}, e_l \in \{1,2,...n\}$), indicating that $\mathbf{x}_l$ takes $\mathbf{x}_{e_l}$ as the cluster exemplar. Data from the same cluster share the same cluster exemplar. We define $\mathbf{e} \triangleq \{e_l\}_{l=1}^n$.
- The quality of the hypothetical labeling $\hat{\mathbf{y}}$ is evaluated by the generalization error bound for the nonparametric plug-in classifier trained by $S_{\hat{\mathbf{y}}}$, and the hypothetical labeling $\hat{\mathbf{y}}$ with minimum associated error bound is preferred, i.e. $\arg\min_{\hat{\mathbf{y}}}\hat{R}_n(\mathrm{PI}_S) = \arg\min_{\hat{\mathbf{y}}}\sum_{l,m}\theta_{lm}G_{lm,\sqrt{2}h_n}$ where $\theta_{lm} = \mathbb{I}_{\hat{y}_l\neq\hat{y}_m}$.
- To avoid the trivial clustering where all the data are grouped into a single cluster, we use the sum of within-cluster dissimilarities term $\sum_{l=1}^n \exp\left(-G_{le_l,\sqrt{2}h_n}\right)$ to control the size of clusters. Therefore, the objective function of our pairwise clustering method is below:

$$\Psi(e) = \sum_{l=1}^n \exp\left(-G_{le_l,\sqrt{2}h_n}\right) + \lambda\sum_{l,m}\left(\tilde{\theta}_{lm}G_{lm,\sqrt{2}h_n} + \rho_{lm}(e_l,e_m)\right)$$

where ρ_{lm} is a function to enforce the consistency of the cluster indicators:

$$\rho_{lm}(e_l,e_m) = \begin{cases}\infty & e_m = l, e_l \neq l \text{ or } e_l = m, e_m \neq m\\ 0 & \text{otherwise}\end{cases},$$

- We show the clustering result on several UCI data sets below.

Data sets	Iris	VC	BT
AP	0.8933 ± 0.0138 (16)	**0.6677** (14)	0.4906 (1)
CEB	0.6929 ± 0.0168 (15)	0.4748 ± 0.0014 (5)	0.3868 ± 0.08 (2)
PIEC	**0.9089** ± **0.0033** (15)	0.5263 ± 0.0173 (35)	**0.6585** ± 0.0103 (5)