Multiclass Total Variation Clustering

Xavier Bresson[1], Thomas Laurent[2], David Uminsky[3] and James von Brecht [4]

[1] University of Lausanne [2] Loyola Marymount University [3] University of San Francisco [4] University of California, Los Angeles

Neural Information Processing Systems Foundation

Summary

Many clustering models rely on the minimization of an energy over possible partitions of the data set. These discrete optimizations usually pose NP-hard problems, however. A natural resolution of this issue involves relaxing the discrete minimization space into a continuous one to obtain an easier minimization procedure. Many current algorithms, such as spectral clustering methods or non-negative matrix factorization (NMF) methods, follow this relaxation approach. Ideas from the image processing literature have recently motivated a new set of algorithms that can obtain better relaxations than those used by NMF and spectral clustering. These new algorithms all rely on the concept of total variation. While these algorithms perform well for bi-partitioning tasks, their recursive extensions yield unimpressive results for multiclass clustering tasks. We present the first general framework for multiclass total variation clustering that does not rely on recursion. Our approach also easily adapts to handle either unsupervised or transductive clustering tasks, and the results significantly outperform previous total variation algorithms and compare well against state-of-the-art approaches

Proximal Splitting Algorithm

We introduce a proximal splitting algorithm to minimize a sum of ratios of convex functions subject to the convex constraint $x \in K$:

$$\min_{x \in K} \sum_{r=1}^{R} \frac{f_r(x)}{g_r(x)} \qquad \text{where } f_r(x) \text{ and } g_r(x) \text{ are convex.}$$

Algorithm:

$$\boxed{\begin{aligned} y^k &= x^k + \partial G^k(x^k) \\ x^{k+1} &= \text{prox}_{F^k + \delta_K}(y^k) \end{aligned}}$$

where $F^k(x)$ and $G^k(x)$ stands for

$$F^k(x) := \tau^k \sum_{r=1}^{R} \left(\frac{1}{g_r(x^k)} \right) f_r(x) \qquad \text{and} \qquad G^k(x) := \tau^k \sum_{r=1}^{R} \left(\frac{f_r(x^k)}{g_r(x^k)^2} \right) g_r(x)$$

The time setp is denoted by τ^k

Energy Estimate:

$$\sum_{r=1}^{R} \frac{g_j(x^{k+1})}{g_j(x^k)} \left(\frac{f_r(x^{k+1})}{g_r(x^{k+1})} - \frac{f_r(x^k)}{g_r(x^k)} \right) \geq \frac{\|x^{k+1} - x^k\|^2}{\tau^k}$$

Inner Stopping Criteria:

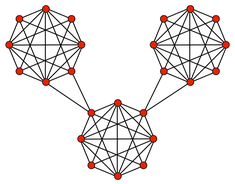
$$\sum_{r=1}^{R} \frac{g_j(x^{k+1})}{g_j(x^k)} \left(\frac{f_r(x^{k+1})}{g_r(x^{k+1})} - \frac{f_r(x^k)}{g_r(x^k)} \right) \geq \theta \frac{\|x^{k+1} - x^k\|^2}{\tau^k}$$

Definition of Proximal Operator:

$$\text{prox}_{F^k + \delta_K}(y^k) := \underset{x \in K}{\text{argmin}} \; F^k(x) + \frac{1}{2} ||x - y^k||^2$$

Asymmetric Balanced Cut

Find the subset A of the data points which minimize



$$\frac{\text{Cut}(A, A^c)}{\min\{\lambda|A|, |A^c|\}}$$

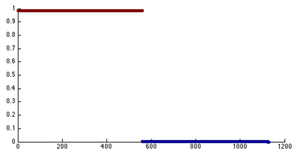
Choosing $\lambda = 2$ will aim at extracting a group containing 1/3 of data points.

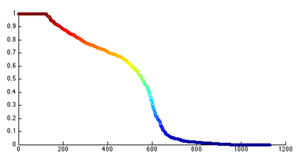
Extract a group containing roughly 1/3 of the data points while cutting as few links as possible.

Multiclass Total Variation Clustering

Total Variation Clustering	Spectral / NMF Clustering
Minimize $\sum_{r=1}^{R} \frac{\text{Cut}(A_r, A_r^c)}{\min\{\lambda|A_r|, |A_r^c|\}}$	Minimize $\sum_{r=1}^{R} \frac{\text{Cut}(A_r, A_r^c)}{|A_r|}$
over all partitions $A_1 \cup \cdots \cup A_R$ of the data points.	over all partitions $A_1 \cup \cdots \cup A_R$ of the data points.
$\Updownarrow$	$\Updownarrow$
Minimize $\sum_{r=1}^{R} \frac{\|f_r\|_{TV}}{\|f_r - \text{med}_\lambda(f_r)\|_{1,\lambda}}$	Minimize $\sum_{r=1}^{R} \frac{f^t L f}{\|f_r\|_2^2}$
over functions $f_1, \ldots, f_R : X \to \{0,1\}$ satisfying $f_1 + \ldots + f_R = \mathbf{1}$.	over functions $f_1, \ldots, f_R : X \to \{0,1\}$ which are mutually orthogonal
$\Downarrow$	$\Downarrow$
Minimize $\sum_{r=1}^{R} \frac{\|f_r\|_{TV}}{\|f_r - \text{med}_\lambda(f_r)\|_{1,\lambda}}$	Minimize $\sum_{r=1}^{R} \frac{f^t L f}{\|f_r\|_2^2}$
over functions $f_1, \ldots, f_R : X \to [0,1]$ satisfying $f_1 + \ldots + f_R = \mathbf{1}$.	over functions $f_1, \ldots, f_R : X \to [0,1]$ which are mutually orthogonal
$\Downarrow$	$\Downarrow$

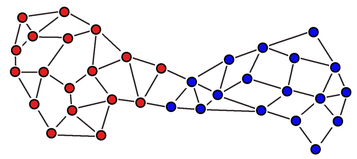
continuous relaxation



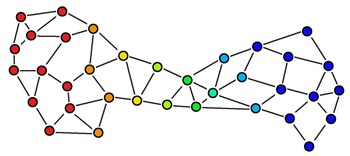


Total Variation and Binary Functions

$$\|f\|_{TV} = \sum_{i,j=1}^{n} w_{ij} |f(\mathbf{x}_i) - f(\mathbf{x}_j)| \qquad f^t L f = \sum_{i,j=1}^{n} w_{ij} |f(\mathbf{x}_i) - f(\mathbf{x}_j)|^2$$



$\|f\|_{TV} = 2$
$f^t L f = 2$



$\|f\|_{TV} = 4.14$
$f^t L f = 0.59$

Experimental results

Unsupervised:

Alg/Data	WEBKB4	OPTDIGITS	PENDIGITS	20NEWS	MNIST
NCC-TV [1]	51.76	95.91	73.25	23.20	88.80
1SPEC [2]	39.68	88.65	82.42	11.49	88.17
LSD [3]	54.50	97.94	88.44	41.25	95.67
NMFR [4]	64.32	97.92	91.21	63.93	96.99
MTV	59.15	98.29	89.06	39.40	97.60

Transductive:

Labels	WEBKB4	OPTDIGITS	PENDIGITS	20NEWS	MNIST
1	56.58/ 1.8s	98.29/ 7s	89.17/ 14s	50.07/ 52s	97.53/ 98s
1%	58.75/ 2.0s	98.29/ 4s	93.73/ 9s	61.70/ 54s	97.59/ 54s
2.5%	57.01/ 1.7s	98.35/ 3s	95.83/ 7s	67.61/ 42s	97.72/ 39s
5%	58.34/ 1.3s	98.38/ 2s	97.98/ 5s	70.51/ 32s	97.79/ 31s
10%	62.01/ 1.2s	98.45/ 2s	98.22/ 4s	73.97/ 25s	98.05/ 25s

[1] X. Bresson, T. Laurent, D. Uminsky and J. H. von Brecht. *NIPS*, 2012.
[2] M. Hein and T. Bühler. *NIPS*, 2010.
[3] R. Arora, M. Gupta, A. Kapila and M. Fazel. *ICML*, 2011.
[4] Z. Yang, T. Hao, O. Dikmen, X. Chen and E. Oja. *NIPS*, 2012.

Acknowledgements

This work was supported by NSF grant DMS-1109805, AFOSR MURI grant FA9550-10-1-0569, ONR grant N000141210040, and Swiss National Science Foundation grant SNSF-141283.