Self-paced Consensus Clustering with Bipartite Graph

Peng Zhou(周芃)[1]; Liang Du(杜亮)[2]; Xuejun Li(李学俊)[1,2]
[1]Anhui University, [2]Shanxi University

Introduction

Clustering often suffers from stable and robust problems.

To tackle this problem, consensus clustering a.k.a. clustering ensemble is proposed.

Consensus clustering aims to learn a consensus clustering result from multiple base clustering results, e.g.

Sample	cluster1	cluster2	cluster3
1	1	0	0
2	1	0	0
3	1	0	0
4	0	1	0
5	0	1	0
6	0	0	1
7	0	0	1
8	0	0	1

Sample	cluster1	cluster2	cluster3
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	1
5	0	0	1
6	1	0	0
7	1	0	0
8	1	0	0

Sample	cluster1	cluster2	cluster3
1	1	0	0
2	1	0	0
3	1	0	0
4	0	1	0
5	0	0	1
6	0	0	1
7	0	0	1
8	0	0	1

Sample	cluster1	cluster2	cluster3
1	1	0	0
2	1	0	0
3	1	0	0
4	0	1	0
5	0	0	1
6	0	0	1
7	0	0	1
8	1	0	0

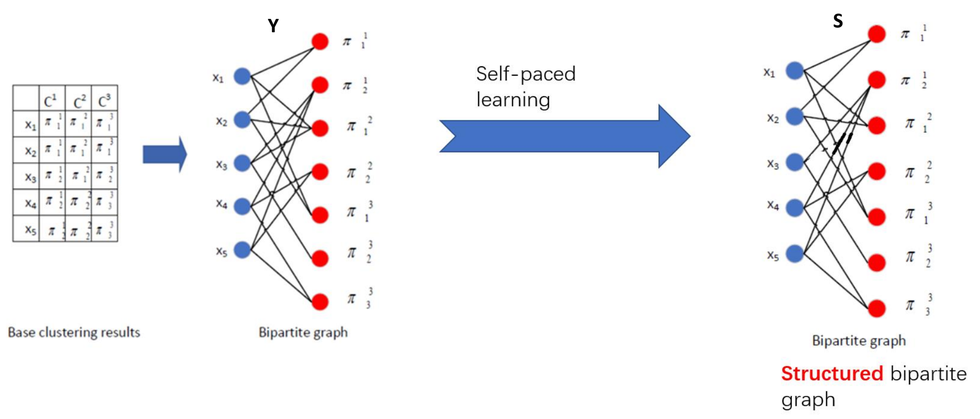
Sample	cluster1	cluster2	cluster3	cluster4
1	1	0	0	0
2	1	0	0	0
3	1	0	0	0
4	0	0	0	1
5	0	1	0	0
6	0	0	1	0
7	0	0	1	0
8	0	0	1	0

Consensus clustering result

Conventional clustering ensemble methods use ***all*** instances for learning, which may be inappropriate, because some instances are ***unreliable***.

To address this issue, we integrate the consensus clustering into ***self-paced learning*** framework, which gradually involve instances from more reliable to less reliable ones into the ensemble learning.

Methods



Y: Initial bipartite graph
S: learned structured bipartite graph
W: reliability matrix of edges
C: similarity matrix of clusters

Objective function:

$$\min_{\mathbf{S},\mathbf{W}} \|\mathbf{W}\odot(\mathbf{S}-\mathbf{Y})\|_F^2 - \lambda\|\mathbf{W}\|_1 + \gamma\sum_{i=1}^{n}\sum_{p,q=1}^{k} C_{pq}(S_{ip}-S_{iq})^2 W_{ip}W_{iq}$$
$$s.t.\quad 0\le S_{ij}\le 1,\quad rank(\mathbf{L}) = n+k-c,$$
$$0\le W_{ij}\le 1.$$

Self-paced regularized term

Learn the reliability of each edge

Make the learned graph contain just c connective components

Optimization:

$$\min_{\mathbf{S},\mathbf{W}} \|\mathbf{W}\odot(\mathbf{S}-\mathbf{Y})\|_F^2 - \lambda\|\mathbf{W}\|_1 + \gamma\sum_{i=1}^{n}\sum_{p,q=1}^{k} C_{pq}(S_{ip}-S_{iq})^2 W_{ip}W_{iq}$$
$$s.t.\quad 0\le S_{ij}\le 1,\quad rank(\mathbf{L}) = n+k-c,$$
$$0\le W_{ij}\le 1.$$

$$\Rightarrow$$

$$\min_{\mathbf{S},\mathbf{W},\mathbf{F}} \|\mathbf{W}\odot(\mathbf{S}-\mathbf{Y})\|_F^2 - \lambda\|\mathbf{W}\|_1 + \gamma\sum_{i=1}^{n}\sum_{p,q=1}^{k} C_{pq}(S_{ip}-S_{iq})^2 W_{ip}W_{iq} + \rho tr(\mathbf{F}^T\mathbf{L}\mathbf{F})$$
$$s.t.\quad 0\le S_{ij}\le 1,\quad 0\le W_{ij}\le 1,\quad \mathbf{F}^T\mathbf{F}=\mathbf{I}.$$

Block coordinate descent algorithm: iteratively update W, S, Y, and F

Results

Methods	ALLAML	GLIOMA	K1b	Lung	Medical	Tr41	Tdt2	Tox
KM	0.6545	0.4239	0.6726	0.7114	0.3996	0.5626	0.4104	0.4229
KM-best	0.7292	0.4880	0.8559	0.8675	0.4707	0.6946	0.4460	0.4825
CSPA	0.6583	0.4100	0.4531	0.4138	0.3500	0.5213	0.2850	0.4246
HGPA	0.5444	0.4180	0.5326	0.5025	0.2950	0.4894	0.2959	0.3854
MCLA	0.6722	0.4000	0.7383	0.7084	0.4017	0.5698	0.4000	0.4152
NMFC	0.6722	0.4140	0.5860	0.6764	0.3789	0.6323	0.3716	0.4269
BCE	0.6708	0.4280	0.6345	0.6700	0.3965	0.6205	0.1806	0.4140
RCE	0.6708	0.4260	0.6887	0.7143	0.3851	0.6391	-	0.4105
MEC	0.6056	0.3940	0.8190	0.7379	0.3627	0.6559	-	**0.4304**
LWEA	0.6736	0.4320	0.8279	0.7458	0.4208	0.6719	0.5744	0.4234
LWGP	0.6750	0.4320	0.7172	0.6498	0.4047	0.6483	0.4288	0.4193
RSEC	0.5917	0.4180	0.8409	0.8217	0.3490	0.6367	0.4222	0.4041
DREC	**0.6819**	0.4280	0.6462	0.6379	0.3926	0.6243	0.3684	0.4205
SCCBG-W	0.6681	0.4080	0.8405	0.8094	0.3980	0.6136	0.5011	0.4053
SCCBG	**0.6861**	**0.4500**	**0.8663**	**0.8961**	**0.4592**	**0.6973**	**0.7164**	**0.4339**

Table 2: ACC results on all the data sets

Methods	ALLAML	GLIOMA	K1b	Lung	Medical	Tr41	Tdt2	Tox
KM	0.0882	0.1629	0.5493	0.5284	0.4209	0.5843	0.6111	0.1374
KM-best	0.1772	0.2347	0.6853	0.6558	0.4806	0.6713	0.6240	0.2164
CSPA	0.0815	0.1716	0.4071	0.3712	0.3992	0.5919	0.5589	0.1436
HGPA	0.0110	0.1509	0.3917	0.3372	0.3613	0.5084	0.5385	0.1083
MCLA	0.0909	0.1327	0.5944	0.5258	0.4296	0.6044	0.6070	0.1329
NMFC	0.0909	0.1550	0.4995	0.5202	0.4259	0.6512	0.5930	0.1434
BCE	0.0821	0.1658	0.5414	0.4977	0.4499	0.6398	0.0000	0.1370
RCE	0.0899	0.1624	0.6068	0.5248	0.4475	0.6499	-	0.1344
MEC	0.0485	0.1312	0.6818	0.5617	0.4089	**0.6758**	-	0.1313
LWEA	0.0935	0.1686	0.6948	0.5364	0.4185	0.6666	0.7183	0.1236
LWGP	0.0932	0.1682	0.6115	0.4993	0.4266	0.6535	0.6266	0.1333
RSEC	0.0495	0.1544	0.6615	0.6027	0.4036	0.6449	0.5243	0.1184
DREC	0.1006	0.1641	0.5774	0.4647	**0.4510**	0.6514	0.5971	0.1394
SCCBG-W	0.0894	0.1567	0.6888	0.5785	0.3220	0.6039	0.6433	0.1239
SCCBG	**0.1252**	**0.2163**	**0.7262**	**0.6930**	0.3918	**0.6847**	**0.7568**	**0.2131**

Table 3: NMI results on all the data sets

Contact

More information (e.g. paper and codes) can be found in Peng Zhou's homepage: https://doctor-nobody.github.io/