IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Heterogeneous Graph Based Similarity Measure for Categorical Data Unsupervised Learning

YANQING YE[1,2], JIANG JIANG*[1,3], BINGFENG GE [1], KEWEI YANG [1], AND H. EUGENE STANLEY[2]

[1]College of Systems Engineering, National University of Defense Technology, Changsha, Hunan 410073, China (e-mail: {yeyanqing09, jiangjiangnudt, bfge.nudt, kayyang27}@nudt.edu.cn)
[2]Center for Polymer Studies, Department of Physics, Boston University, Boston, MA 02215, U.S.A. (e-mail:{yanqing, hes}@bu.edu)
[3]Channing Division of Network Medicine, Harvard Medical School, Boston, MA 02115, USA.

*Corresponding author: Jiang Jiang (e-mail: jiangjiangnudt@nudt.edu.cn).

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 71671186 and 71690233, and the Research Project of National University of Defense Technology.

I. SUPPLEMENTARY MATERIALS

As the limit of the scope, we present the spectral and k-modes clustering results evaluated by purity and rand index and the related analysis here.

A. COMPARISON OF HGS WITH OTHER SIMILARITY MEASURES DERIVED SPECTRAL CLUSTERING BY PURITY AND RAND INDEX

Table 1 and Table 2 respectively show the results evaluated by purity and rand index. From the average measure, our HGS method obtains the best average *RI* with 0.609 and ranks second according to purity. while the following one is Lin method, which has performed best according to purity, however, it only defeats OF method when evaluated by *RI*. CMS ranks second by *RI*, however, it's inferior to Lin, our HGS, and Hamming distance according to purity. More precisely, from the perspective of *RI*, HGS outperforms other measures in 8 datasets, while according to *F-score*, HGS achieves best in 5 datasets. Similarly, CMS wins in 8 datasets via *RI* and 4 datasets via purity. On the contrary, OF measure gets the worst results on the average for both *RI* and purity, which are 0.579 and 0.653, respectively. No measure outperforms all other measures in all datasets. Every measure has performed best in at least one dataset. Besides, the methods that capture the co-occurrence of the attribute values perform better on the whole. Therefore, it's essential to explore the relationships between attribute values when measuring the similarity of objects.

B. COMPARISON OF HGS WITH OTHER SIMILARITY MEASURES BASED K-MODES CLUSTERING BY RAND INDEX AND PURITY

The evaluation results and the average measure of *RI* and purity for k-modes clustering were shown in Table 3 and Table 4, respectively. Combining both evaluations, our HGS and Hamming distance perform an equally excellent performance, where HGS performs the best by purity with 0.725 while Hamming is best for *RI* with 0.614. Both measures rank second according to another metric. Subsequently, CMS ranks third according to both metrics. Again, OF gets the worst results according to both metrics. More accurately, from the perspective of *RI*, HGS and ALGO respectively perform the best in 9 datasets while Hamming is best in 4 datasets. According to purity, HGS has outperformed other methods in 11 out of 26 datasets while ALGO wins in 7 datasets. Subsequently, Lin wins in 6 datasets via purity and 4 datasets by *RI*. Therefore, from the performance on the whole and the detailed dataset, we can conclude that our proposed HGS method can perform better than other methods in the k-modes clustering task.

• • •

TABLE 1: The Rand Index (*RI*) of Hamming, OF, Lin, ALGO, CMS vs. HGS-based spectral clustering

Dataset	Hamming	OF	Lin	ALGO	CMS	HGS
Sos	0.818	0.577	**1.000**	0.834	0.818	0.834
Ha	0.621	0.591	0.618	0.566	**0.631**	0.618
He	0.523	0.525	0.529	0.535	0.526	**0.553**
Br	0.848	0.711	0.597	0.742	**0.851**	0.785
Ho	0.593	0.594	0.567	**0.690**	0.603	0.657
Sol	0.791	0.629	**0.809**	0.723	0.773	0.813
SP	0.402	0.400	0.369	0.375	0.401	**0.442**
Zo	0.950	0.918	0.891	0.942	0.913	**0.963**
DN	0.572	0.615	0.566	0.516	0.568	**0.645**
Ly	0.553	0.557	0.544	**0.571**	0.543	0.537
Mo	0.519	0.522	**0.528**	0.501	0.521	0.513
De	0.767	0.720	**0.867**	0.859	0.834	0.746
Cr	0.556	0.553	0.541	0.520	**0.568**	0.562
Ma	0.606	**0.622**	0.596	0.616	0.605	0.603
Fl	0.402	0.575	0.366	0.390	0.432	**0.439**
Pr	0.700	0.647	0.805	**0.815**	0.632	0.604
Ti	0.488	0.462	0.464	0.472	**0.495**	0.489
Ba	0.568	0.575	0.575	0.561	**0.578**	0.577
Ca	0.463	0.463	0.463	0.464	0.463	**0.475**
Ch	0.500	0.502	**0.503**	0.504	0.500	0.500
Cw	0.515	0.520	0.509	0.522	0.523	**0.541**
Im	0.563	0.452	0.444	0.371	**0.608**	0.501
Ip	**0.595**	**0.595**	0.586	0.484	**0.595**	**0.595**
Or	0.521	0.500	0.500	0.501	**0.528**	0.527
Bs	0.522	0.613	0.506	**0.834**	0.639	0.654
Bss	0.522	0.613	0.506	**0.834**	0.639	0.654
Average	0.595	0.579	0.586	0.605	0.607	**0.609**

TABLE 2: The Purity of Hamming, OF, Lin, ALGO, CMS vs. HGS-based spectral clustering

Dataset	Hamming	OF	Lin	ALGO	CMS	HGS
Sos	0.766	0.574	**1.000**	0.787	0.766	0.787
Ha	0.485	0.553	0.485	**0.561**	0.553	0.523
He	0.676	0.690	**0.725**	0.697	0.676	0.697
Br	0.965	0.849	0.944	0.956	**0.971**	0.909
Ho	0.884	0.884	0.866	**0.935**	0.884	0.909
Sol	0.417	0.357	**0.489**	0.368	0.444	0.451
SP	0.794	0.794	0.794	**0.813**	0.794	0.794
Zo	**0.911**	0.822	0.842	0.851	0.822	0.891
DN	0.783	0.811	0.774	0.623	0.755	**0.858**
Ly	0.635	0.628	0.635	**0.723**	0.601	0.588
Mo	0.728	0.759	**0.757**	0.561	0.725	0.685
De	0.686	0.639	0.730	**0.762**	0.760	0.656
Cr	**0.814**	0.781	0.806	0.702	**0.814**	0.803
Ma	0.823	0.816	0.828	**0.837**	0.831	0.824
Fl	**0.836**	0.835	0.829	0.834	0.835	**0.836**
Pr	0.333	0.295	**0.402**	0.371	0.288	0.280
Ti	0.864	0.686	0.720	0.661	**0.899**	0.883
Ba	0.616	0.669	**0.699**	0.538	0.685	0.670
Ca	0.702	0.700	0.701	0.700	0.706	**0.811**
Ch	0.552	0.595	0.592	**0.612**	0.540	0.554
Cw	0.653	**0.673**	0.653	0.663	0.643	**0.673**
Im	**0.913**	0.897	0.889	0.857	0.889	**0.913**
Ip	0.756	0.756	0.756	**0.767**	0.756	0.756
Or	0.635	0.635	0.635	0.635	**0.651**	0.635
Bs	0.206	0.133	0.193	**0.305**	0.135	0.149
Bss	0.206	0.133	0.193	**0.305**	0.135	0.149
Average	0.678	0.653	**0.690**	0.670	0.675	0.680

TABLE 3: The Rand Index of Hamming, OF, Lin, ALGO, CMS vs. HGS-enabled k-modes Clustering

Dataset	Hamming	OF	Lin	ALGO	CMS	HGS
Sos	**1.000**	0.883	0.805	0.870	0.827	0.828
Ha	0.611	0.645	0.636	0.616	0.636	**0.652**
He	0.515	0.514	0.523	0.520	0.520	**0.528**
Br	**0.545**	0.528	0.509	0.538	0.523	0.540
Ho	**0.591**	0.587	0.580	0.589	0.582	0.589
Sol	0.891	0.883	**0.901**	0.880	0.892	0.889
SP	0.369	0.367	0.367	**0.393**	0.367	0.363
Zo	0.848	0.856	0.839	0.847	0.854	**0.858**
DN	0.538	0.532	0.539	**0.613**	0.544	0.542
Ly	0.588	0.563	0.567	0.557	0.560	**0.594**
Mo	0.503	0.507	0.504	0.506	**0.508**	0.503
De	0.860	0.782	**0.883**	0.881	0.870	0.873
Cr	0.526	**0.530**	0.521	0.520	**0.530**	0.529
Ma	0.578	0.580	0.552	0.565	0.581	**0.584**
Fl	0.340	**0.337**	0.334	**0.337**	0.336	0.335
Pr	0.809	0.792	0.832	**0.845**	0.810	0.826
Ti	0.464	0.462	0.462	0.463	0.465	**0.466**
Ba	0.563	0.562	0.563	0.430	0.565	**0.571**
Ca	0.463	0.463	0.463	**0.542**	0.463	0.468
Ch	0.506	0.506	0.507	**0.510**	0.507	0.507
Cw	0.506	0.507	0.505	**0.512**	0.505	**0.512**
Im	0.458	0.338	0.330	0.334	**0.503**	0.419
Ip	0.485	0.479	0.487	0.460	0.479	**0.522**
Or	**0.505**	0.495	0.490	0.496	0.501	0.495
Bs	0.951	0.937	**0.953**	**0.953**	0.941	0.949
Bss	0.951	0.937	**0.953**	**0.953**	0.941	0.949
Average	**0.614**	0.599	0.600	0.605	0.608	0.611

TABLE 4: The Purity of Hamming, OF, Lin, ALGO, CMS vs. HGS-enabled k-modes Clustering

Dataset	Hamming	OF	Lin	ALGO	CMS	HGS
Sos	**1.000**	0.872	0.766	0.851	0.787	0.787
Ha	0.508	0.644	0.606	0.576	0.598	**0.659**
He	0.697	0.669	0.718	0.704	0.697	**0.732**
Br	0.958	0.921	0.949	0.958	**0.965**	0.946
Ho	0.914	0.909	0.909	**0.918**	0.909	0.914
Sol	0.575	0.549	**0.586**	0.553	**0.586**	0.571
SP	0.794	0.794	0.794	**0.824**	0.794	0.794
Zo	**0.832**	**0.832**	0.812	**0.832**	0.822	**0.832**
DN	0.717	0.689	0.745	**0.811**	0.726	0.726
Ly	**0.757**	0.743	0.696	0.709	0.696	**0.757**
Mo	0.612	0.651	0.629	0.647	**0.662**	0.599
De	0.855	0.628	**0.934**	0.932	0.896	0.896
Cr	0.797	0.823	0.751	0.739	0.820	**0.836**
Ma	0.829	0.804	0.808	0.824	0.816	**0.832**
Fl	0.829	0.830	0.829	**0.831**	0.830	0.829
Pr	**0.432**	0.386	**0.432**	0.424	0.402	**0.432**
Ti	0.691	0.676	0.700	0.678	0.717	**0.729**
Ba	0.573	0.568	0.573	0.461	0.587	**0.618**
Ca	0.709	0.700	0.709	0.700	0.701	**0.742**
Ch	0.714	0.721	**0.723**	0.699	0.718	0.719
Cw	0.643	0.643	0.643	**0.663**	0.633	0.653
Im	0.865	0.857	0.857	0.857	**0.889**	0.865
Ip	0.733	0.744	0.767	0.733	0.744	**0.778**
Or	0.619	0.619	0.619	**0.635**	0.619	0.603
Bs	0.502	0.488	**0.504**	0.500	0.487	0.501
Bss	0.502	0.488	**0.504**	0.500	0.487	0.501
Average	0.718	0.702	0.714	0.714	0.715	**0.725**