Self-Supervised Clustering based on Manifold Learning and Graph Convolutional Networks

1. Supplementary Material

This document presents the supplementary contents, composed of tables that present more details about conducted experiments not included in the original paper due to space limitations. The remaining of this document is organized as follows: Section 2 discusses our parameter selection approach comparing results presented in the original paper. Section 3 presents an ablation study that evaluates the impact of different steps of the method by removing or replacing components of the proposed architecture with classical methods in the literature.

2. Parameter Selection

As stated in our conclusion, a more in-depth parameter estimation study will be conducted in future work. However, we have suggested values for parameters k (size of the neighborhood explored by LHRR) and t (number of interactions executed by LHRR) in the experiments presented in the paper. This section explores the presented results, discusses the effectiveness of our approach, and gives some insights on how to select parameters for specific use cases.

First, we compare the results obtained by our proposed approach with a fixed value of $k = 50$ with methods from literature (listed in Tables 1 and 3 from the original paper, respectively). Regarding image datasets, the results for $k = 50$ are the best in all but two specific metrics: the Normalized Mutual Information (NMI) in both CUB200 and MPEG-7 datasets. Although, in both scenarios our approach obtained better metrics in V-Measure (VM) and, specially, in Accuracy (ACC). Regarding citation networks, SGCC obtained the best results in ACC for Cora and in VM for Citeseer. Therefore, our method presented SOTA or comparable results in almost every evaluated scenario with our suggested parameter.

In a second analysis, we explore the results from parameter t following the procedure conducted for parameter p in Section 4.2 of the original paper. For this experiment, we defined $k = 50$ and $p = 0.5$ while varying t between 1 and 2. However, the results were not presented in detail due to space limitations. Table 1 presents the experiment results on the Corel5K dataset. The best results for all networks were obtained with $t = 1$. However, the difference between the evaluated metrics is small in all cases, close to 1%. Moreover, both scenarios ($t = 1$ and $t = 2$) presented competitive results when compared with methods from literature which indicate the robustness of our method to different parameter t definitions.

Table 1. Comparison of the results for different values of the parameter t in the Corel5k dataset. All experiments were conducted with parameters $k = 50$ and $p = 0.5$.

Network	t	NMI	V-Measure	ACC
GCN	1	**91.34 ± 00.16**	**91.10 ± 00.15**	**88.45 ± 00.12**
	2	90.60 ± 00.10	90.07 ± 00.11	87.54 ± 00.13
SGCC	1	**91.74 ± 00.06**	**91.50 ± 00.06**	**88.74 ± 00.05**
	2	91.27 ± 00.03	90.76 ± 00.04	88.15 ± 00.04
APPNP	1	**91.64 ± 00.15**	**90.47 ± 00.15**	**88.72 ± 00.15**
	2	90.97 ± 00.14	90.46 ± 00.15	88.02 ± 00.16

3. Ablation Study

The construction of our approach is composed of three main steps: *(i)* The *Log-based Hypergraph Ranking of References (LHRR)* [3] manifold learning method explores the dataset and provide meaningful structures; *(ii)* The LHHR's structures are used to identify representative elements and to cluster a portion of the data, producing reliable soft-labels for training a *Graph Neural Network (GCN)*; *(iii)* The trained GCN is used to define the final cluster labels for the complete dataset.

This section explores the removal or replacing of these three components in the clustering task. In this scenario, for better understanding of the experiments, Table 2 presents the nomenclature used to identify which components are being used or removed in the result tables present in the following sub-sections.

Table 2. Nomenclature used in the experiments' result tables, for better understanding.

Nomenclature	Description
confid	Computation of the Hyperedge Self-confidence Score from structures obtained from LHRR to define the elements clusterized in the second step.
rep	Use of the selected representative elements in the clustering step executed to obtain soft-labels for the GCN's training procedure.
rl	Use of the improved ranked lists obtained from LHRR in the GCN's training procedure.
gcn	Training and application of a GCN in order to cluster the remaining dataset samples.

The experiments are organized as follows. Section 3.1 explores the effects of removing LHRR as the starting point for our approach. Section 3.2 presents the results obtained by replacing our initial clustering step, which defines soft-labels for the final training procedure, by *K-Means*. Finally, Section 3.3 analyzes the results of not using a GCN for the approach's final step.

3.1. Removing LHRR

In this first experiment, LHRR was removed from the proposed approach. In this scenario, we need to understand and replace four functionalities provided by the manifold learning method: *(i)* The Hyperedge Self-Confidence Score (described in Section 3.2.1 from the original paper), which defines the dataset's order for selecting the slice of the dataset used for the creation of soft-labels; *(ii)* The Representatives Proxy Selection (described in Section 3.2.2 of the original paper), which drives the initial clusterization of the select elements from the dataset; *(iii)* The clustering approach, which uses the hyperedges from LHRR's hypergraph; and *(iv)* The k-NN graph used during the training of the GCN model, for image datasets, which is created using the improved ranked lists obtained from LHRR.

As a substitute for the Hyperedge Self-confidence Score, we used the Authority Score [1], which explores the ranked list for each element o_i in the dataset measuring how many elements are shared between the explored ranked list and the ranked lists of each element present in its first k positions. As a result, reliable elements obtain higher scores since they are able to contain other similar elements in the first position of their ranked lists. The Authority Score is used to order the dataset's elements, selecting which elements will be clusterized and used as soft-labels.

Furthermore, we remove the selection of representative elements, which requires a different metric capable of measuring the similarity between two elements (as presented on Equation 10 from the original publication), and apply K-Means [2] as the clustering approach for defining the soft-labels. Finally, the k-NN graph used to train the GCN model is computed based on the ranked lists obtained from the original features.

During experimentation, we apply this new configuration in two image datasets: Corel5K and Flowers. Additionally, following the experimental protocol from the original paper, parameters k and t are varied in the same intervals: 10 and 100 for k; 1 and 2 for t. Furthermore, half of the dataset is classified as soft-labels ($p = 0.5$) and the *SGC* GCN network is used in the experiments. Finally, reported values are the mean and standard deviation over 10 executions.

Table 3 presents the comparison between the configuration proposed in this section, which removes LHRR's execution, and the best results obtained by the proposed approach in both image datasets cited above. SGCC obtained the best results in all evaluated metrics in both datasets. This experiment highlighted the importance of the reliable information provided by LHRR's hypergraph, which is explored in all steps of our proposed approach.

3.2. Replace Our Clustering Approach

During the second step of our proposed approach, after the selection of representative elements from the dataset, a clustering agglomeration is conducted, exploring the hyperedges provided by LHRR, to define soft-labels for the semi-supervised training of the GCNs. In this experiment, we replace this clustering step by the *K-means* and compare the results obtained in two image datasets with the best results reported for SGCC in the original paper.

Regarding the components described in Table 2, in this experiment we explore two different scenarios: *(i)* The selected portion of the dataset is clustered by K-means directly; and *(ii)* Alongside the selected elements, we use our selected representatives (rep) as clusters centers in K-means initialization. The remaining of the experimental protocol is configured as explained in Section 3.1: The configurations are evaluated in Corel5K and Flowers datasets, parameters k e t are explored in their defined intervals, the SGC model is applied to conduct the final separation of the dataset, and the mean and standard deviation values after 10 iterations are reported as the final result.

Therefore, we compare both configurations with the best results from SGCC using the SGC model for fair comparison. Table 4 presents the experiment results. As K-Means achieved effective results in the image datasets with competitive results in Table 3 from the original paper, the results from both datasets in both configurations are high in all metrics. Furthermore, the results obtained by the configuration that used the representatives as centers in the Corel5K dataset are better than every other literature method evaluated in the original paper. However, the results are still below the values obtained by SGCC.

Additionally, the best k for Corel5K goes down from 100 to 70 when the representatives are used in K-Means, being equal to the best configuration of SGCC. This behavior highlights the impact of the representatives selected, as described in Section 3.2.2 of the original paper, as the initial definition for each of the expected clusters.

Finally, the difference of results between the configurations using K-Means and the SGCC is higher in Flowers, probably due to the fact of the Flowers dataset ResNet features produce ranked lists not so reliable as the ones produced by Corel5K's ResNet features. In this scenario, the errors of the original features are minimized by SGCC's clustering step by using the structures obtained from LHRR's hypergraph.

Table 3. Comparison between our approach and a new configuration, which removes the LHRR manifold learning method and applies an Authority Score to define which elements to clusterize and the K-Means algorithm as clustering method.

Dataset	Method	confid	rep	rl	gcn	k	t	NMI	V-Measure	ACC
Corel5K	**Authority + K-Means**	×	×	×	✓	10	1	88.22 ± 00.67	86.85 ± 00.70	75.27 ± 01.83
	SGCC	✓	✓	✓	✓	70	2	$\mathbf{92.62 \pm 00.06}$	$\mathbf{92.44 \pm 00.06}$	$\mathbf{90.80 \pm 00.04}$
Flowers	**Authority + K-Means**	×	×	×	✓	20	2	75.84 ± 01.07	74.65 ± 01.13	69.55 ± 02.84
	SGCC	✓	✓	✓	✓	50	2	$\mathbf{81.27 \pm 00.07}$	$\mathbf{81.01 \pm 00.07}$	$\mathbf{83.49 \pm 00.09}$

Table 4. Evaluation of the replacement of SGCC's clustering step with K-Means in two different configurations: with and without the selected representatives (rep) used as clusters centers during K-Means initialization.

Dataset	Method	confid	rep	rl	gcn	k	t	NMI	V-Measure	ACC
Corel5K	**LHRR + K-Means**	✓	×	✓	✓	100	2	91.82 ± 00.29	90.75 ± 00.26	84.21 ± 01.30
	LHRR + K-Means	✓	✓	✓	✓	70	2	92.11 ± 00.04	91.95 ± 00.04	90.70 ± 00.02
	SGCC	✓	✓	✓	✓	70	2	$\mathbf{92.62 \pm 00.06}$	$\mathbf{92.44 \pm 00.06}$	$\mathbf{90.80 \pm 00.04}$
Flowers	**LHRR + K-Means**	✓	×	✓	✓	60	2	77.84 ± 00.70	77.14 ± 00.82	74.53 ± 02.26
	LHRR + K-Means	✓	✓	✓	✓	60	2	79.32 ± 00.16	78.95 ± 00.15	79.00 ± 00.12
	SGCC	✓	✓	✓	✓	50	2	$\mathbf{81.27 \pm 00.07}$	$\mathbf{81.01 \pm 00.07}$	$\mathbf{83.49 \pm 00.09}$

Table 5. Comparison of two version of the SGCC methods, with and without the final GCN step to separate the dataset's elements.

Dataset	Method	confid	rep	rl	gcn	k	t	NMI	V-Measure	ACC
Corel5K	**SGCC - no GCN**	✓	✓	✓	×	50	1	92.47	92.24	89.62
	SGCC (SGC)	✓	✓	✓	✓	70	2	$\mathbf{92.62 \pm 00.06}$	$\mathbf{92.44 \pm 00.06}$	$\mathbf{90.80 \pm 00.04}$
Flowers	**SGCC - no GCN**	✓	✓	✓	×	35	1	**81.88**	**81.83**	**83.75**
	SGCC (SGC)	✓	✓	✓	✓	50	2	81.27 ± 00.07	81.01 ± 00.07	83.49 ± 00.09
Cora	**SGCC - no GCN**	✓	✓	✓	×	95	1	31.76	31.13	53.24
	SGCC (SGC)	✓	✓	✓	✓	50	2	$\mathbf{45.02 \pm 00.15}$	$\mathbf{44.81 \pm 00.15}$	$\mathbf{62.96 \pm 00.09}$

3.3. Removing the GCN

As a final experiment, we explore the results obtained by clustering all dataset elements in the second step of our proposed approach, described in Section 3.2.3 from the original paper, using the n_c function, presented in Equation 12 from the original paper, and setting parameter $p = 1$. In this scenario, the final step involving the training of the GCN model is not used, since every element is already separated in the earlier step. Using this configuration, SGCC becomes a deterministic method, extending the deterministic nature of LHRR.

This section presents the comparison of a configuration that removes the final GCN step compared with the best results from SGCC. Following the other experiments from this ablation study, parameters k and t are varied inside their respective intervals. Additionally, every step before the GCN training is maintained. Finally, the comparison was conducted over three datasets: Corel5K, Flowers, Cora.

Table 5 presents the results from this experiment. The results from the configuration presented in this section, without the GCN step, achieved great results in both images datasets. Besides not achieving higher values in the Corel5K dataset, it was able to surpass the proposed method in the Flowers dataset.

However, the results from the original model are better in two of the three experiments and have obtained metrics almost matching the results from this configuration in the Flowers dataset, with differences smaller than 1%.

Additionally, the difference in applying the GCN training is highlighted in the Cora dataset experiment. The possibility to use graphs, especially external graphs that can provide different information about the data, allows SGCC to obtain almost 10% increase in every evaluated metric.

Moreover, another indirect proof of the GCN training for the separation of data can be seen in Table 4, where the results of K-Means separation after applying the GCN step were higher than the obtained by K-Means directly, in the original paper.

Finally, we believe that the possibility to change which GCN model is applied to our approach also opens a series of possibilities since new models are still being proposed in the literature and can be easily integrated with SGCC.

References

[1] Daniel Carlos Guimarães Pedronette, Otávio A.B. Penatti, and Ricardo da S. Torres. Unsupervised manifold learning using reciprocal knn graphs in image re-ranking and rank aggregation tasks. *Image and Vision Computing*, 32(2):120–130, 2014. 2

[2] James B. MacQueen. Some methods for classification and analysis of multivariate observations. 1967. 2

[3] Daniel Carlos Guimarães Pedronette, Lucas Pascotti Valem, Jurandy Almeida, and Ricardo da S. Torres. Multimedia retrieval through unsupervised hypergraph-based manifold ranking. *IEEE Transactions on Image Processing*, 28(12):5824–5838, 2019. 1