Large-scale Spectral Clustering with Stochastic Nyström Approximation

Hongjie Jia[1], Liangjun Wang[1] and Heping Song[1]

[1] Jiangsu University, Zhenjiang Jiangsu 212013, China
jiahj@ujs.edu.cn

Abstract. In spectral clustering, Nyström approximation is a powerful technique to reduce the time and space cost of matrix decomposition. However, in order to ensure the accurate approximation, a sufficient number of samples are needed. In very large datasets, the internal singular value decomposition (SVD) of Nyström will also spend a large amount of calculation and almost impossible. To solve this problem, this paper proposes a large-scale spectral clustering algorithm with stochastic Nyström approximation. This algorithm uses the stochastic low rank matrix approximation technique to decompose the sampled sub-matrix within the Nyström procedure, losing a slight of accuracy in exchange for a significant improvement of the algorithm efficiency. The performance of the proposed algorithm is tested on benchmark data sets and the clustering results demonstrate its effectiveness.

Keywords: Spectral Clustering, Nyström Approximation, Stochastic SVD, Large Dataset.

1 Introduction

Spectral clustering algorithms can well deal with the datasets of non-convex structures, and they have been successfully applied in many fields. But the traditional spectral clustering algorithms only suit for small-scale datasets, because they need to store an $n \times n$ affinity matrix and make eigen-decomposition on it. The required space complexity and time complexity are respectively $O(n^2)$ and $O(n^3)$. The high complexity problems limit the application of spectral clustering methods in large data [1]. Therefore it is needed to develop a new data processing strategy to adapt to the continuous growth of the data size while maintaining the quality and speed of the clustering.

Fortunately, the spectral clustering method only needs a small part of the head (or tail) of the eigenvalues / eigenvectors, then we can use the Arnoldi method to do partial SVD [2]. However, experience shows that only when the matrix is sparse or very few eigenvectors are extracted, the running time will be significantly reduced. Another method to reduce the computational complexity is using low rank matrix approximation, such as the commonly used Nyström method [3]. It selects a subset of $m << n$ columns from the kernel matrix, and then constructs the low rank approximation of the kernel matrix by using the correlation between the samples and the remaining

columns. In computation, the Nyström method only requires the decomposition of a small $m \times m$ sub-matrix and the time complexity can be significantly reduced; in the occupied space, it is only need to store the sampled m columns, and other matrix involved in the computation can be calculated by the m columns, so its space complexity is small. This makes the Nyström method has high scalability. Fowlkes et al. [4] successfully apply it to spectral clustering for image segmentation. In order to improve the accuracy of Nyström approximation, we need to select a lot of samples, but the large sampled sub-matrix is also very difficult to decompose [5].

Halko et al. [6] propose a stochastic SVD method to construct approximate low rank matrix factorization. This method extends the Monte Carlo algorithm in literature [7]. Similar to the standard Nyström method, this method only need the eigen-decomposition on part of matrix. But this method does not simply select a subset of columns, it first construct a low dimensional subspace that captures the activity of the input matrix. Then, compress the matrix into the subspace, and make the standard factorization on the reduced matrix. Although the method is a stochastic algorithm, experiments show that it has great potential to produce accurate approximations. However, this algorithm needs to traverse at least once the input matrix, so it is more time-consuming than the Nyström method which is only based on sampled columns. On large data sets, the performance difference between these algorithms will be significant.

In this paper, we combine the advantages of standard Nyström method and stochastic SVD algorithm. Standard Nyström is very efficient, but need to collect a large number of columns; stochastic SVD algorithm has high accuracy, but the efficiency is relatively low. Inspired by this, when using the Nyström method for large scale spectral clustering, we can use stochastic SVD to replace the original standard SVD on the sampled sub-matrix to cope with efficiency decrease caused by the increasing sample number m, and accelerate the process of calculating approximate eigenvectors. The main contributions of this paper are:

- We propose a large-scale spectral clustering algorithm with stochastic Nyström approximation, which can achieve a good balance between the clustering accuracy and the operating efficiency.
- The approximation error of stochastic SVD process in the proposed method can be compensated by selecting more sample columns.
- Experimental results show that the proposed method can further reduce the calculation complexity of Nyström spectral clustering.

The rest of this paper is organized as follows. Section 2 briefly reviews the related research background. Section 3 introduces the proposed large-scale spectral clustering algorithm with stochastic Nyström approximation. The experimental results are given in Section 4, and the last section is conclusion.

2 Research Background

2.1 Nyström Approximation

The spectral methods such as Ratio Cut and Normalized Cut are based on the eigenvectors of Laplacian matrix to do clustering [8]. Suppose the Laplacian matrix is $L = D^{-1/2}WD^{-1/2}$, where D is the degree matrix and W is the weight matrix. The eigenvector matrix U can be calculated by the eigen-decomposition of Laplacian matrix L, namely $LU = U\Lambda_L$. The eigenvectors in matrix U are orthogonal and these eigenvectors embed the data objects into a low dimensional subspace. Then we may use k-means algorithm to cluster U, and obtain the final partition results. When the amount of data n is very large, it becomes very difficult to decompose the Laplacian matrix. The Nyström method use a subset of matrix columns (or rows) to do approximate spectral decomposition for a large matrix, which can significantly reduce the computational complexity [9].

Given the $n \times n$ weight matrix W, we randomly select $m << n$ data points from data set and rearrange matrix W as follows:

$$W = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \tag{1}$$

where $A \in \mathbb{R}^{m \times m}$ contains the similarities among the data samples, $B \in \mathbb{R}^{m \times (n-m)}$ contains the similarities among the samples and the rest points, and $C \in \mathbb{R}^{(n-m) \times (n-m)}$ contains the similarities among the rest points.

Nyström approximation gets the approximate eigenvectors of Laplacian matrix using the eigenvectors of a small sub-matrix. Let matrix $H = \begin{bmatrix} A & B \end{bmatrix}^T$, matrix W can be approximated as:

$$\tilde{W} = HA^{-1}H^T \tag{2}$$

To get the orthogonal approximate eigenvectors of $\tilde{W}$, Fowlkes et al. [4] define a matrix $M = A + A^{-1/2}BB^TA^{-1/2}$. We decompose M as $M = U_M \Lambda_M U_M^T$. The eigenvector matrix of $\tilde{W}$ are:

$$U_W = HA^{-1/2}U_M\Lambda_M^{-1/2} \tag{3}$$

where $\tilde{W} = U_W \Lambda_M U_W^T$. It can be proved that U_W and its transpose matrix are orthogonal, namely $U_W^T U_W = I$.

In order to compute the first k approximate eigenvectors and eigenvalues of W, the total time complexity of this algorithm is $O(m^3 + kmn)$, where $O(m^3)$ is the eigen-decomposition time of M, and $O(kmn)$ is corresponding to the multiply operations about matrix H. Because $m << n$, its complexity is much lower than the $O(n^3)$ complexity that directly SVD on W. Although the efficiency of Nyström method is high, it needs to

select a sufficient number of samples to better approximate the original eigen-space. Then we consider use stochastic SVD to further reduce the complexity of Nyström method.

2.2 Stochastic SVD

Halko et al. [6] propose a simple and efficient stochastic SVD algorithm, which is used to solve the approximate eigenvalues and eigenvectors of the low rank matrix. Given a real symmetric matrix $M \in \mathbb{R}^{m \times m}$, this stochastic SVD algorithm includes two stages: first, it uses random sampling to construct a low dimensional subspace to approximate the range of M; then, it limits M in the obtained sub-space, and makes standard QR or SVD decomposition based on the reduced matrix. The following algorithm 1 gives the concrete steps of the stochastic SVD, through which we can quickly obtain a low rank approximation of a real symmetric matrix M.

Algorithm 1. Stochastic SVD.

Input: symmetric matrix $M \in \mathbb{R}^{m \times m}$, matrix rank k, over sampling parameter p, power parameter q.

Output: the eigenvector matrix U_M, the eigenvalue matrix Λ_F.

Step 1. Construct an $m \times (k + p)$ standard Gaussian random matrix Ω.

Step 2. Calculate matrix $Z = M\Omega$ and $Y = M^{q-1}Z$.

Step 3. Find an orthogonal matrix Q through the QR decomposition, so that $Y = QQ^TY$.

Step 4. According to $F(Q^T\Omega) = Q^TZ$, compute the matrix F.

Step 5. Conduct SVD on F and get $F = U_F \Lambda_F U_F^T$.

Step 6. Compute matrix $U_M = QU_F$.

Specifically, the first stage of Algorithm 1 includes Step 1 ~ Step 3. It first produces an $m \times (k + p)$ standard Gaussian random matrix Ω, each element of Ω is independent Gaussian random variables, the mean is 0, the variance is 1. Among them, p is an over sampling parameter, so that the column number of Ω is slightly higher than the required rank k. Then calculate the matrix $Y = M\Omega$, and construct the matrix $Q \in \mathbb{R}^{m \times (k+p)}$ through the QR decomposition. Q is an orthogonal matrix, and its column constitutes the orthogonal basis of Y. In order to make $Y = M\Omega$ have a larger range to extend to the k dimensional subspace of M, the value of p is generally a small number, such as 5 or 10.

The second stage of Algorithm 1 includes Step 4 ~ Step 6. M is restricted to the subspace generated by Y, we can further obtain the reduced matrix $F = Q^TMQ$. And then conduct the standard SVD on F, that is $F = U_F \Lambda_F U_F^T$. The SVD of M can be approximated as:

$$M \simeq QFQ^T = (QU_F)\Lambda_F (QU_F)^T \tag{4}$$

Finally, let $U_M = QU_F$, we can obtain the low rank approximation of M as $M \approx U_M \Lambda_F U_M^T$. The time complexity of Algorithm 1 is $O(m^2k + k^3)$, which is proportional to the square of m. Algorithm 1 is easy to implement, and can be applied

to large scale clustering problem. Therefore, we introduce the stochastic SVD into Nyström approximation to deal with the complex eigen-decomposition problem when the sampled sub-matrix is too large.

3 Large-Scale Spectral Clustering with Stochastic Nyström Approximation

The Nyström approximation technology uses the sample points to compute the approximate eigenvectors. It can effectively reduce the computational complexity of traditional spectral clustering algorithm. The performance of Nyström approximation is closely related to sample number. Although improving the sampling proportion can improve the clustering results, the complexity of the algorithm is also significantly increased. Careful observation can be found that when the sample number m is large, the most time-consuming operation of the algorithm is the eigen-decomposition of the $m \times m$ sub-matrix M. In order to solve this problem, we develop the stochastic Nyström approximation method to solve the approximate eigenvalue and eigenvector of M. We try to improve the efficiency of the algorithm as far as possible in the premise of ensuring the clustering accuracy. Therefore we propose a large-scale spectral clustering algorithm with stochastic Nyström approximation. The details of the proposed algorithm is shown in Algorithm 2.

Algorithm 2. Large-scale spectral clustering with stochastic Nyström approximation (SNA-SC).

Input: data set X of n data points, number of sample points m, number of classes k.

Output: clustering results of k clusters.

Step 1. According to Equation (1), form matrix $A \in \mathbb{R}^{m \times m}$ and matrix $B \in \mathbb{R}^{m \times (n-m)}$ with the m sample points.

Step 2. Calculate the diagonal degree matrix $D = \mathrm{diag}\left(\begin{bmatrix} A\mathbf{1}_m + B\mathbf{1}_{n-m} \\ B^T\mathbf{1}_m + B^T A^{-1} B\mathbf{1}_{n-m} \end{bmatrix} \right)$ with matrix A and B.

Step 3. Calculate the normalized matrix A and B as $\overline{A} = D_{1:m,1:m}^{-1/2} A D_{1:m,1:m}^{-1/2}$, $\overline{B} = D_{1:m,1:m}^{-1/2} B D_{m+1:n,m+1:n}^{-1/2}$, and form matrix $H = \begin{bmatrix} \overline{A} & \overline{B} \end{bmatrix}^T$.

Step 4. Construct matrix $M = \overline{A} + \overline{A}^{-1/2} \overline{B}\overline{B}^T \overline{A}^{-1/2}$ with matrix $\overline{A}$ and $\overline{B}$.

Step 5. Make the eigen-decomposition of M by Algorithm 1, namely $M \approx U_M \Lambda_F U_M^T$, and ensure the descending order of the eigenvalues in Λ_F.

Step 6. Calculate the top k orthogonal eigenvectors of the Laplacian matrix using Equation (3): $\tilde{V} = H\overline{A}^{-1/2} (U_M)_{:,1:k} (\Lambda_F^{-1/2})_{1:k,1:k}$.

Step 7. Normalize matrix $\tilde{V}$ by Equation (10) and get matrix $\tilde{U}$.

$$\tilde{U}_{ij} = \frac{\tilde{V}_{ij}}{\sqrt{\sum_{r=1}^{k} \tilde{V}_{ir}^2}}, \quad i = 1, \cdots, n, \quad j = 1, \cdots, k \tag{5}$$

Step 8. The rows of $\tilde{U}$ can be seen as new data points and we can divide them into k clusters by traditional clustering algorithms, such as k-means.

The proposed Algorithm 2 combines the advantages of Nyström approximation and the stochastic SVD, and has a good performance in the clustering efficiency and accuracy. In essence, the low rank approximation of the original $n \times n$ affinity matrix W can be expressed as $\tilde{W} = HA^{-1}H^T = U_W \Lambda_M U_W^T = U_W U_M^T M U_M U_W^T$ according to Equation (3). Through Algorithm 1 and Equation (4), we can obtain the approximate M as $M \simeq QFQ^T$. So the more accurate approximation form of W in Algorithm 2 is as follows:

$$\tilde{W} \simeq U_W U_M^T Q F Q^T U_M U_W^T \tag{6}$$

Different with the Nyström method, Algorithm 2 adopts stochastic SVD method for solving the approximate eigenvalues and eigenvectors of matrix M. Its time complexity is $O(m^2k + k^3)$. In addition, the matrix H related multiplication operations need to spend $O(kmn)$ time. Usually $n >> m \geq k$, so the total time complexity of Algorithm 2 is $O(k^3 + kmn)$. Compared to the $O(m^3 + kmn)$ complexity of Nyström method, the complexity of Algorithm 2 is lower. This means that, for the same size of problem, Algorithm 2 can finish the task in a shorter time.

4 Experimental Analysis

To validate the performance of the proposed SNA-SC algorithm, our experiments are done on the four real world data sets from UCI machine learning repository. These data sets are listed in Table 1.

Table 1. Basic properties of the data sets.

Data set	Data points' number	Attributes' number	Clusters' number
Corel	2074	144	18
Seismic	98528	50	3
RCV1	193844	47236	103

Based on the data sets in Table 1, we compare three different clustering algorithms in the experiments. In addition to the proposed SNA-SC algorithm, there are approximate kernel k-means algorithm (AKK-means) [10], the spectral clustering algorithm based on Nyström extension (Nyström-SC) [11]. The performance of each algorithm are evaluated by the clustering accuracy and running time. All algorithms are implemented by MATLAB, running on a high-performance workstation with 3.20GHz CPU. In the experiments, the affinities of data points are measured by radial basis function. The

max iterations of AKK-means algorithm is 1000. The sample points in Nyström-SC and SNA-SC algorithm are obtained by random sampling.

Table 2. Clustering accuracy of algorithms (%).

Data set	Sampling ratio (m/n)	Algorithm		
		AKK-means	Nyström-SC	SNA-SC
Corel	2%	31.62 (±1.51)	30.42 (±0.76)	31.21 (±1.63)
	4%	34.16 (±1.07)	30.92 (±0.85)	33.64 (±1.41)
	6%	37.27 (±0.54)	32.46 (±0.72)	36.86 (±1.25)
	8%	**38.58** (±0.76)	33.27 (±0.57)	37.16 (±1.08)
Seismic	2%	62.46 (±1.24)	60.48 (±1.56)	64.13 (±1.29)
	4%	64.14 (±0.43)	62.57 (±1.12)	67.23 (±0.42)
	6%	64.81 (±0.62)	63.44 (±1.81)	67.75 (±0.83)
	8%	65.27 (±0.53)	64.91 (±0.34)	**68.34** (±0.37)
RCV1	2%	12.55 (±0.76)	11.27 (±0.65)	12.43 (±0.84)
	4%	13.82 (±0.53)	14.24 (±0.46)	14.67 (±0.39)
	6%	15.42 (±0.72)	16.73 (±0.67)	15.94 (±0.65)
	8%	16.26 (±0.51)	**18.12** (±0.34)	16.63 (±0.41)

Table 2 is the clustering accuracy of these algorithms on each data set, in which the bold value is the best clustering result. AKK-means, Nyström-SC and SNA-SC use part of the kernel matrix for approximate computation, so they need to sample some data points. From Table 2, we know that the clustering accuracy of each algorithm is different in different sampling proportion. Overall, the increase in the proportion of sampling is helpful to improve the quality of clustering. However, sometimes more samples will also make the clustering quality slightly worse, because it contains more noise data. AKK-means algorithm constructs the approximate kernel matrix by random sampling, and based on this, it can reduce the space complexity of the original kernel *k*-means by computing the class center of kernel *k*-means in a smaller subspace. AKK-means clustering can get the highest accuracy on Corel data set. However, on other data sets, AKK-means is not as good as Nyström-SC and SNA-SC algorithm. Nyström-SC is suitable for processing RCV1 data set. SNA-SC has good performance on Seismic data set.

The clustering time of different algorithms are compared in Table 3. Table 3 shows that SNA-SC has the highest running efficiency on each data set. On RCV1 data set, SNA-SC only takes 44 seconds to do the clustering under 8% sampling rate, while Nyström-SC takes 162 seconds. Because with the sampling rate increase, the decomposition of the internal sub-matrix in Nyström-SC will cost a lot of time. AKK-means is a k-means-like algorithm. It repeatedly relocate the cluster center to optimize the lose function. The clustering time of AKK-means is mainly related to the iteration times. Although more samples will help increase the approximation accuracy, but it also increases the clustering time. For AKK-means and SNA-SC, their clustering time

increase linearly with the sampling ratio increasing. But the clustering time of Nyström-SC has violent changes because of the cubic time complexity of the eigen-decomposition of the internal sub-matrix.

Table 3. Clustering time of algorithms (s).

Data set	Sampling ratio (m/n)	Algorithm		
		AKK-means	Nyström-SC	SNA-SC
Corel	2%	0.11	0.06	0.05
	4%	0.14	0.09	0.05
	6%	0.16	0.10	0.07
	8%	0.18	0.36	0.08
Seismic	2%	6.10	0.82	0.72
	4%	7.69	3.03	1.75
	6%	11.68	6.15	2.68
	8%	19.38	26.75	5.86
RCV1	2%	40.44	21.36	20.66
	4%	56.01	27.33	25.02
	6%	73.94	64.29	31.25
	8%	110.96	162.32	44.12

5 Conclusion

Nyström approximation will help reduce the complexity of spectral clustering using approximate eigenvectors. However, when the sample number is too large, internal SVD of Nyström will take a very long time. This paper applies the stochastic SVD algorithm to improve the performance of large-scale Nyström spectral clustering. Unlike standard Nyström method, we use the stochastic low rank matrix approximation strategy to do the eigen-decomposition of the internal sub-matrix, and propose a large-scale spectral clustering called SNA-SC. Experimental results show that SNA-SC is more efficient than standard Nyström spectral clustering, and it can well balance the clustering accuracy and efficiency.

Acknowledgement. This work was supported by the National Natural Science Foundations of China (grant numbers 61906077, 61601202), the Natural Science Foundation of Jiangsu Province (grant numbers BK20190838, BK20170558), and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (grant number 18KJB520009, 16KJB520008).

References

1. Kang, Z., Shi, G., Huang, S., Chen, W., Pu, X., Zhou, J. T., Xu, Z.: Multi-graph fusion for multi-view spectral clustering. Knowledge-Based Systems 189, 105102 (2020).
2. Tang, M., Marin, D., Ayed, I. B., Boykov, Y.: Kernel cuts: Kernel and spectral clustering meet regularization. International Journal of Computer Vision 127(5), 477-511 (2019).
3. Jia, H., Ding, S., Du, M.: A Nyström spectral clustering algorithm based on probability incremental sampling. Soft Computing 21(19), 5815-5827 (2017).
4. Fowlkes, C., Belongie, S., Chung, F., Malik, J.: Spectral grouping using the Nystrom method. IEEE transactions on pattern analysis and machine intelligence 26(2), 214-225 (2004).
5. Li, M., Bi, W., Kwok, J. T., Lu, B. L.: Large-scale Nyström kernel matrix approximation using randomized SVD. IEEE transactions on neural networks and learning systems 26(1), 152-164 (2014).
6. Halko, N., Martinsson, P. G., Tropp, J. A.: Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM review 53(2), 217-288 (2011).
7. Drineas, P., Kannan, R., Mahoney, M. W.: Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix. SIAM Journal on computing 36(1), 158-183 (2006).
8. Jia, H., Ding, S., Du, M., Xue, Y.: Approximate normalized cuts without Eigen-decomposition. Information Sciences 374, 135-150 (2016).
9. Wang, S., Gittens, A., Mahoney, M. W.: Scalable kernel K-means clustering with Nyström approximation: relative-error bounds. The Journal of Machine Learning Research 20(1), 431-479 (2019).
10. Chitta, R., Jin, R., Havens, T. C., Jain, A. K.: Approximate kernel k-means: Solution to large scale kernel clustering. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 895-903. ACM, San Diego (2011)..
11. Chen, W. Y., Song, Y., Bai, H., Lin, C. J., Chang, E. Y.: Parallel spectral clustering in distributed systems. IEEE transactions on pattern analysis and machine intelligence 33(3), 568-586 (2011).