A Randomized Approach to Efficient Kernel Clustering

Grebenkin Ivan

Abstract—**Kernel-based K-means is widely used but has drawbacks such as the proportionality of memory consumption to the square of points in the sample. The authors of the article "A Randomized Approach to Efficient Kernel Clustering" provide a new analysis of a class of approximate kernel methods that have more modest memory requirements, and propose a specific one-pass randomized kernel approximation followed by standard Kmeans on the transformed data. This result was verified in this paper.**

Index Terms: Kernel methods, Unsupervised learning, Low-rank approximation, Randomized algorithm

I. Introduction

Kernel machines have been widely used in various machine learning problems such as classification, clustering, and regression. In kernel-based learning, the input data points are mapped to a high-dimensional feature space and the pairwise inner products in the lifted space are computed and stored in a positive semidefinite kernel matrix K. The lifted representation may lead to better performance of the learning problem, but a drawback is the to store and manipulate a large kernel matrix of size $n \times n$, where n is the size of data set. Thus a kernel machine has quadratic space complexity and quadratic or cubic computational complexity (depending on the specific type of machine). One promising strategy for reducing these costs consists of a low-rank approximation of the kernel matrix $\mathbf{K} = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L} \in \mathbb{R}^{n \times r}$ for a rank $r < n$. Such low-rank approximations can be used to reduce the memory and computation cost by trading-off accuracy for scalability. or this reason, much research has focused on efficient algorithms for computing low-rank approximations.

The Nystroem method works by selecting a small set of bases referred to as landmark points and computing the kernel similarities between the input data points and landmark points. Therefore, the performance of the Nystroem method depends crucially on the number of selected landmark points as well as the procedure according to which these landmark points are selected.

The original one-pass Nystroem method is based on sampling a small subset of m colunms of $\mathbf{K}$ using uniform sampling without replacement. This one-pass algorithm to be true requires at least two passes over the kernel matrix.

Proposed algorithm is not necessarily faster than the Nystroem approach, but has lower memory requirements.

II. Notation and Preliminaries

Column vectors with lower-case bold letters and matrices with upper-case bold letters. For vector $\mathbf{x} \in \mathbb{R}^p$, let $||\mathbf{x}||_2$ be the Euclidean norm, and $diag(x)$ is a diagonal matrix with the elements of $\mathbf{x}$ on the main diagonal. For the matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ the Frobenius norm is $||\mathbf{A}||_F = (\sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij}^2)^{\frac{1}{2}} = (tr(\mathbf{A}^T\mathbf{A}))^{\frac{1}{2}}$, where A_{ij} is the (i,j)-th entry of $\mathbf{A}$,$\mathbf{A}^T$ is a transpose $\mathbf{A}$, $tr(\cdot)$ is a trace operator. $\mathbf{K} \in \mathbb{R}^{n \times n}$ is a symmetric positive semidefinite matrix with $rank(\mathbf{K} = \rho \leq n)$. The singular value decomposition or eigenvalue decomposition of $\mathbf{K}$ is a $\mathbf{K} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$

III. Theory

Let $\mathbf{X} = [\mathbf{x_1}, ..., \mathbf{x_n}] \in \mathbb{R}^{p \times n}$ be a data matrix that contains n data points in $\mathbb{R}^p$. Let the inner products in feature space are calculated using a kernel function $k(\cdot,\cdot)$ defined on the original space:

$$k(\mathbf{x}_i, \mathbf{x}_j) =< \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) >$$

where Φ is the kernel-induced feature map. A well-studied approach to reduce the memory and computation burden associated with kernel machines is to use a low-rank approximation of kernel matrices. This approach utilizes the decaying spectra of kernel matrices and the best rank-r approximation

$$\mathbf{K}_{(r)} = \mathbf{U}_r \mathbf{\Lambda}_r \mathbf{U}_r^T$$

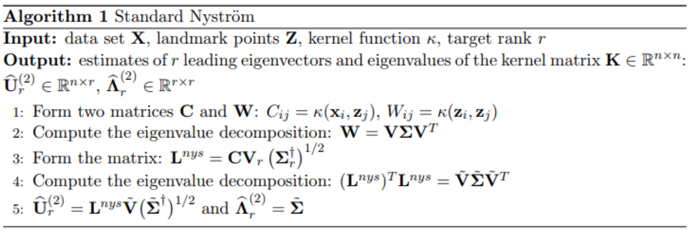
is computed. Since K is symmetric positive semidefinite, the eigenvalue decomposition can be used to express a low-rank approximation in the form of:

$$\mathbf{K}_{(r)} = \mathbf{L}\mathbf{L}^T, \mathbf{L} = \mathbf{U}_r \mathbf{\Lambda}_r^{\frac{1}{2}} \in \mathbb{R}^{n \times r}$$

In this paper the kernels obtained in three different ways, were compared. The first kernel was computed by using the polynomial kernel of order $d = 2$, i.e

$$k(\mathbf{x}_i, \mathbf{x}_j) =< \mathbf{x}_i, \mathbf{x}_j >^2$$

The second one has been obtained by original Nystroem method.



Algorithm 1 Standard Nyström

Input: data set $\mathbf{X}$, landmark points $\mathbf{Z}$, kernel function κ, target rank r

Output: estimates of r leading eigenvectors and eigenvalues of the kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$: $\widehat{\mathbf{U}}_r^{(2)} \in \mathbb{R}^{n \times r}$, $\widehat{\mathbf{\Lambda}}_r^{(2)} \in \mathbb{R}^{r \times r}$

```
Form two matrices C and W: $C_{ij} = \kappa(\mathbf{x}_i, \mathbf{z}_j)$, $W_{ij} = \kappa(\mathbf{z}_i, \mathbf{z}_j)$
Compute the eigenvalue decomposition: $\mathbf{W} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^T$
Form the matrix: $\mathbf{L}^{nys} = \mathbf{C}\mathbf{V}_r \left(\mathbf{\Sigma}_r^{\dagger}\right)^{1/2}$
Compute the eigenvalue decomposition: $(\mathbf{L}^{nys})^T\mathbf{L}^{nys} = \tilde{\mathbf{V}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^T$
$\widehat{\mathbf{U}}_r^{(2)} = \mathbf{L}^{nys}\tilde{\mathbf{V}}(\tilde{\mathbf{\Sigma}}^{\dagger})^{1/2}$ and $\widehat{\mathbf{\Lambda}}_r^{(2)} = \tilde{\mathbf{\Sigma}}$
```

The last one obtained by proposed algorithm.

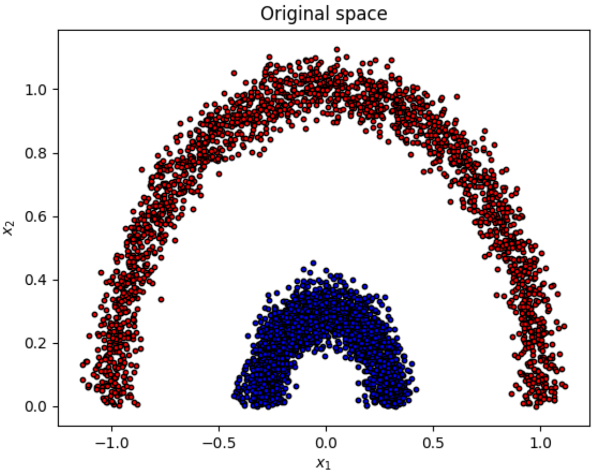
Algorithm 1 One-Pass Kernel K-means

Input: kernel matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$, rank r, oversampling l, number of clusters K, number of iterations
Output: cluster indicator matrix $\mathbf{C}$

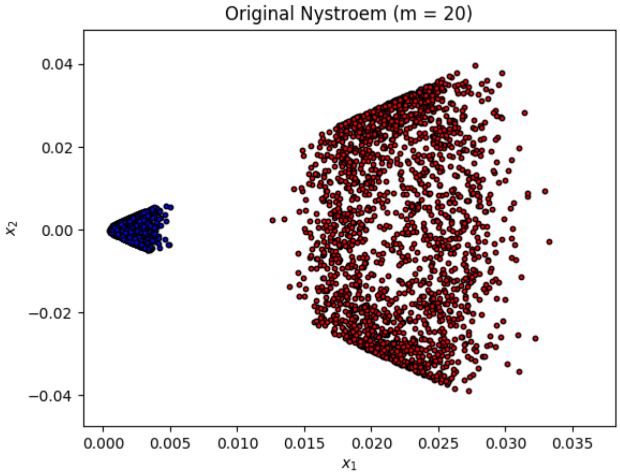
```
1: $r' \leftarrow r + l$,   $\mathbf{R} \in \mathbb{R}^{n\times r'}$: random sampling matrix
2: $\mathbf{W} \in \mathbb{R}^{n\times r'} \leftarrow (\mathbf{R}^T\mathbf{HDK})^T$
3: find an orthonormal matrix $\mathbf{Q} \in \mathbb{R}^{n\times r}$ by QR decomposition or
   $r$ leading left singular vectors of $\mathbf{W}$
4: solve $\mathbf{B}(\mathbf{Q}^T\mathbf{\Omega}) = (\mathbf{Q}^T\mathbf{W})$
5: $\mathbf{B} = \mathbf{V\Sigma V}^T$
6: $\mathbf{Y} = \mathbf{\Sigma}^{1/2}\mathbf{V}^T\mathbf{Q}^T \in \mathbb{R}^{r\times n}$
7: perform standard K-means on $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$ in $\mathbb{R}^r$
```

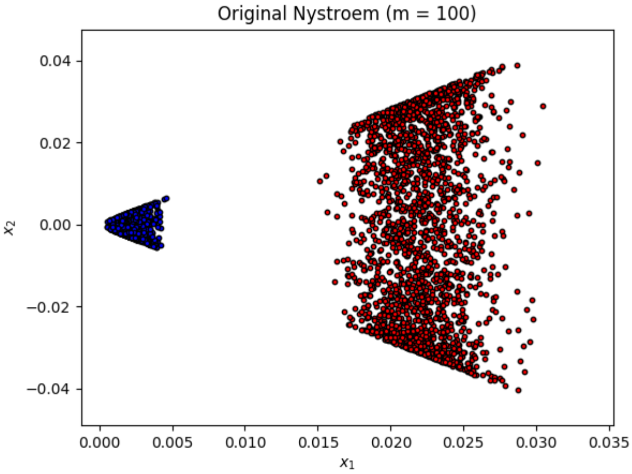
IV. Results

Visualization of mapping results is given below. Original data. As data to verify the operation of the algorithm, synthetic data were used *make_circles* from *sklearn.datasets*.

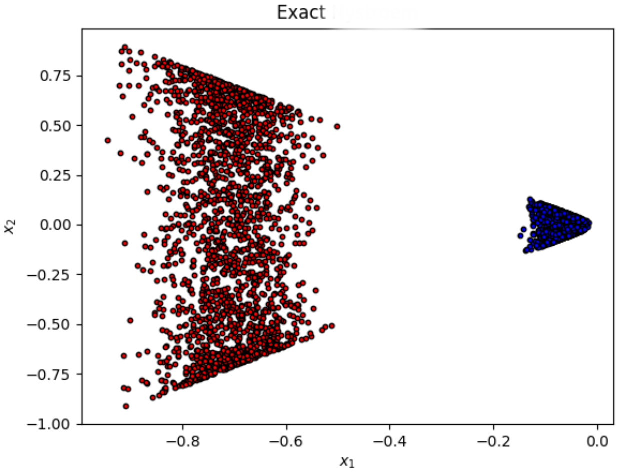


Then, mapping of the original data using the low-rank approximation was made. For this purpose the standard Nystroem was used.

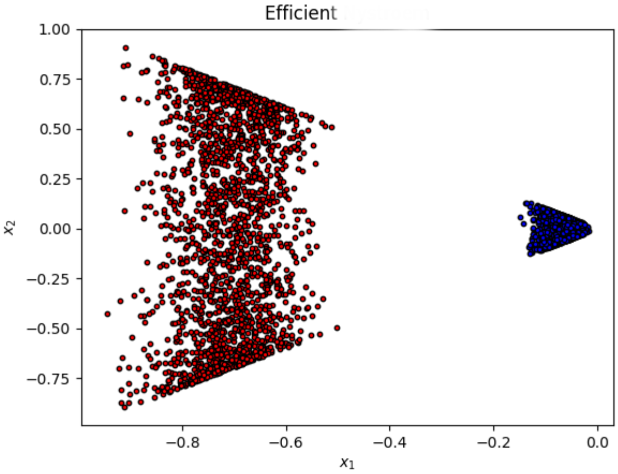




The exact decomposition by using polynomial kernel of order $d = 2$ shown below.



Due to randomized part of algorithms there is rotation of space. The last figure is a proposed efficient algorithm result.

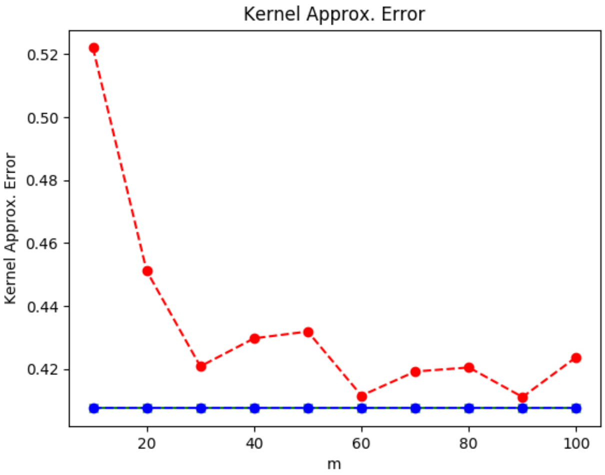


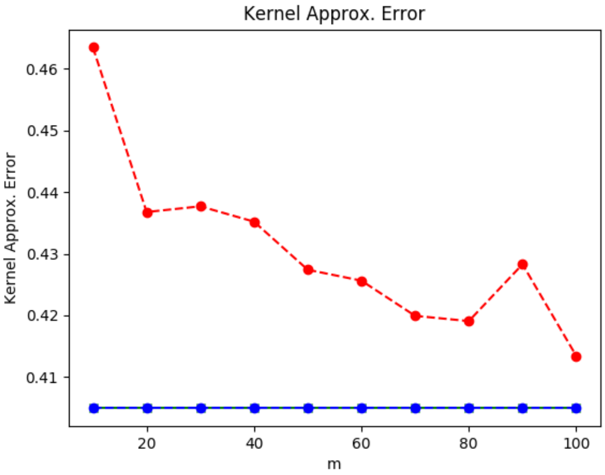
To assess the quality of the proposed algorithm, the Kernel Approximation Error and Clustering Accuracy were compared for these three approaches.

For these purposes, the data set proposed in the original article was used. It is segmentation data set that can be downloaded from the UCI Repository. It has 19 features and 7 clusters.

Figures below show the normalized approximation error of the kernel matrix obtained by using Frobenius norm $||\mathbf{K} - \hat{\mathbf{K}}||_F/||\mathbf{K}||_F$, where $\hat{\mathbf{K}}$ is a obtained with algorithm and $\mathbf{K}$ is an exact decomposition by using polynomial kernel of order $d = 2$.

Red line is a Standart Nystroem. Blue is an efficient algorithm approximation error and green is an exact decomposition. Two last are coincided on the graph.





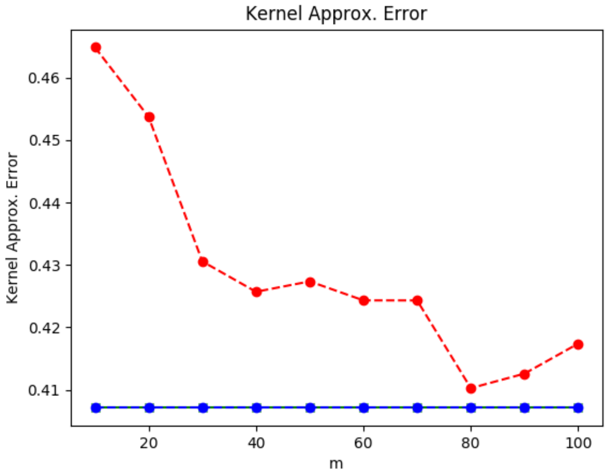
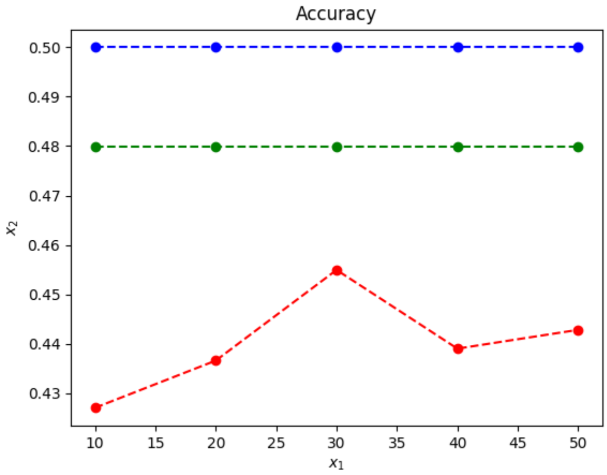


Figure below shows clustering accuracy. Clustering Accuracy was measured for each cluster by cross-checks, because of the randomness of the algorithm, it changes clusters in places, but it still segments clusters correctly.



V. Conclusion

The results of the original article were checked. The original algorithm of Nystroem was implemented. Algorithm also proposed the original article. The obtained graphs show that the proposed algorithm shows the effectiveness close to the effectiveness of a quadratic polynomial kernel. Measurements were made with the same value of memory/speed tradeoff parameter as in the original article. At the same time, the synthetic dataset of *circles* was used for plotting. And segmentation data set from the UCI Repository was used for measurements.

VI. Repository

https://github.com/cubazis/randomized_kernel_clustering

References

[1] Farhad Pourkamali-Anaraki, Stephen Becker. A RANDOMIZED APPROACH TO EFFICIENT KERNEL CLUSTERING. *IEEE GlobalSIP 2016* University of Colorado at Boulder.