

DC^2 : A Divide-and-conquer Algorithm for Large-scale Kernel Learning with Application to Clustering

Ke Alexander Wang^{†*}, Xinran Bian^{‡*}, Pan Liu[¶], Donghui Yan[§]

[†]Cornell University, Ithaca, NY

[‡]Shanghai Jiao Tong University, China

[¶]Zhejiang University, China

[§]University of Massachusetts Dartmouth, MA

November 19, 2019

Abstract

Divide-and-conquer is a general strategy to deal with large scale problems. It is typically applied to generate ensemble instances, which potentially limits the problem size it can handle. Additionally, the data are often divided by random sampling which may be suboptimal. To address these concerns, we propose the DC^2 algorithm. Instead of ensemble instances, we produce structure-preserving signature pieces to be assembled and conquered. DC^2 achieves the efficiency of sampling-based large scale kernel methods while enabling parallel multicore or clustered computation. The data partition and subsequent compression are unified by recursive random projections. Empirically dividing the data by random projections induces smaller mean squared approximation errors than conventional random sampling. The power of DC^2 is demonstrated by our clustering algorithm rpfCluster^+ , which is as accurate as some fastest approximate spectral clustering algorithms while maintaining a running time close to that of K-means clustering. Analysis on DC^2 when applied to spectral clustering shows that the loss in clustering accuracy due to data division and reduction is upper bounded by the data approximation error which would vanish with recursive random projections. Due to its easy implementation and flexibility, we expect DC^2 to be applicable to general large scale learning problems.

Index terms— Divide-and-conquer, large scale kernel learning, clustering, random projection forests, recursive random projections

1 Introduction

Kernel learning is an important problem in machine learning that lies at the core of kernel methods [35, 22]. For example, support vector machines [12], and

*Contributed equally.

various kernelized methods such as kernel PCA [34], kernel ridge regression [33], kernel ICA [1], kernel CCA [19], kernel k-means [14] etc all require the learning of a kernel. Additionally, kernels have been used in semi-supervised learning, for example, [3] uses the graph Laplacian. Furthermore, all spectral clustering algorithms [36, 30, 38, 43] involve the learning of a similarity kernel.

There are several attractive properties with kernels. It allows the embedding of potentially unstructured data to a space suitable for learning and inference, and often effectively overcomes the *curse of dimensionality* arising from high dimensional data. However, as the kernel requires the computing of pairwise similarity for all points, the computational complexity for learning the kernel is very high ($O(n^2)$ for a naive implementation on n points). Indeed many spectral clustering algorithms, e.g., [36, 30], have a computational complexity of $O(n^3)$. A number of algorithms have been proposed to speed up the computation. For example, the Nyström algorithm [17, 15], and some data-dependent sampling algorithms [43, 7]. While such algorithms generally work remarkably well, they do not take advantage of today's widely available multicore or clustered computing infrastructure. We propose a divide-and-conquer approach to split a large-scale kernel learning problem such that subtasks can run in parallel on multicore or clustered computers.

Divide-and-conquer is a popular strategy to deal with large or complex problems when the subproblems are easier to solve or faster to compute. Divide-and-conquer works by dividing a problem into smaller subproblems, working on the subproblems, and then aggregating results from subproblems into the solution to the original problem. Classic divide-and-conquer algorithms include Quicksort [21], the Karatsuba algorithm for multiplying large numbers [24], fast Fourier transform [10] etc. For large scale computation, there are generally two possible ways to implement divide-and-conquer. One is to divide the algorithm into parallel components and conquer on each. However, this is algorithm dependent since it requires tailoring the divide-and-conquer to the algorithm's implementation details. Another approach, which is popular and easy to implement, is to partition the data such that each partition forms a subproblem. As the subproblems use different partitions of the data, they can run independently. This allows one to take advantage of the multicore or clustered computing infrastructure where subproblems are computed in parallel to speed up the overall computation. If the target algorithm has a super-linear computational complexity and the aggregation of subproblems is "easy", one can still achieve remarkable speedup by solving the subproblems *sequentially*, especially when the original computation takes large computing resources. Dividing the data and then conquering each component also makes it possible to tackle a larger problem than forming an ensemble instance on each data partition, as different partitions can be used to learn the target (e.g., data representation) on disjoint supports. In contrast, directly ensembling on the similarity matrix would severely limit scalability as each partition now has to learn the full similarity matrix which is not desirable due to its $O(n^2)$ complexity.

The most straightforward way to divide the data is by random sampling. This is particularly easy to implement and suitable for ensemble-based algorithms. However, random sampling does not account for structure in the data and may

be suboptimal. As a remedy, we propose to use random projections to divide the data. Our method gives more accurate and more robust results by taking into account the geometry of the data. Given the high computational complexity of kernel learning, a further representation compression over each partition is often necessary. Representation compression obtains a structure-preserving signature of the data which can be used in place of the full dataset to speed up computations. We will use recursive random projections [13, 44] to produce a compressed signature for each partition. The idea of recursive random projections has been successfully applied in fast approximate spectral clustering [43], computing over distributed data [45, 46], and other procedures. Since our approach is a divide-and-conquer method with a representation compression, we refer to it as *divide-compress-and-conquer*, or DC^2 in short.

Our main contributions are as follows. First, we propose a geometry-aware divide-compress-and-conquer algorithm for large scale kernel learning. It unifies the division of the data and the subsequent representation compression over each partition through recursive random projections. It incurs negligible approximation error in the resulting kernel, and a vanishing loss in accuracy when applied to clustering. As our approach is based on data partition, it allows parallel computation of subtasks on each data partition with multicore or clustered computing. Our proposed algorithm is easy to implement and readily applies to high dimensional data without the potential curse of dimensionality. Beyond kernel learning, it is immediately applicable to general large scale learning and inference problems.

The remainder of this paper is organized as follows. In Section 2, we give a detailed description of the DC^2 algorithm and its application to a kernel-based clustering algorithm. This is followed by a theoretical analysis on the approximation error of DC^2 and the resulting loss in clustering accuracy. Related work are discussed in Section 4. In Section 5, we empirically evaluate the approximation error by recursive random projections in DC^2 and we compare a DC^2 -enabled clustering algorithm to its competitors. Finally, we conclude in Section 6.

2 Proposed approach

In this section, we will describe the DC^2 algorithm for large scale kernel learning. DC^2 consists of three steps: 1) data partition; 2) representation compression; 3) conquering and aggregation of subtasks. For concreteness of description, we will apply the DC^2 algorithm to *rpfCluster* [42], a clustering algorithm based on a data-driven kernel learned by random projection forests (rpForests) [44]; the resulting algorithm is termed *rpfCluster⁺*. We will describe the DC^2 algorithm and *rpfCluster⁺* in the rest of this section.

2.1 The DC^2 algorithm

We divide the data by recursive random projections. We first project all the data onto a randomly generated direction, then divide the data into two halves by the median of the projections (or by a randomly chosen split point). If more

sively, instead of partitioning all at once along a single random direction, is to avoid thin and long slices of data which will potentially harm the performance of the subsequent clustering [44].

Now we can give an algorithmic description of the DC^2 algorithm. Let V denote the set of all points. Let n_p denote the predefined number of data partitions (e.g., same as the number of cores on the machine), and n_s be the node size below which rpTrees will not grow further. As DC^2 uses rpTrees, we also include a description of rpTrees. The algorithm for DC^2 and rpTrees are termed as Algorithm 2 and Algorithm 1, respectively.

Algorithm 1 $rpTree(D)$

```

1: Let  $D$  be the root node of tree  $t$ ;
2: Initialize the working queue  $\mathcal{W} \leftarrow \{D\}$ ;
3: while  $\mathcal{W}$  is not empty do
4:   Take node  $W$  from  $\mathcal{W}$ ;
5:   If  $|W| < n_s$ , then skip to next round of the loop;
6:   Split node  $W$  by random projection into  $W = W_L \cup W_R$ ;
7:   Add  $W_L, W_R$  to queue  $\mathcal{W}$  and also tree  $t$ ;
8: end while
9: return( $t$ );

```

Algorithm 2 $DC^2(V)$

```

1: Initialize  $S \leftarrow \emptyset, \mathcal{W} \leftarrow \{V\}$ ;
2: for  $i = 1$  to  $n_p$  do
3:   Take the largest node  $W$  from  $\mathcal{W}$ ;
4:   Split node  $W$  by random projection into  $W = W_L \cup W_R$ ;
5:   Add  $W_L, W_R$  to queue  $\mathcal{W}$ ;
6: end for
7: for  $i = 1$  to  $n_p$  do
8:   Take node  $W$  from queue  $\mathcal{W}$ ;
9:   Grow random projection tree  $t_i \leftarrow rpTree(W)$ ;
10:  Let  $S_i$  be the set of leaf node signatures for tree  $t_i$ ;
11:  Update  $S \leftarrow S \cup S_i$ ;
12: end for
13: return( $S$ );

```

2.2 $rpfCluster$

$rpfCluster$ [42] is a clustering algorithm based on the learning of the $rp\text{-kernel}$ via $rp\text{Forests}$ [44]. $rp\text{Forests}$ is an ensemble of random projection trees (rpTrees) [13] with the possibility of projection selection during tree growth. Instead of splitting the nodes along coordinate-aligning axes such as the popular kd-tree [4], rpTrees recursively splits the tree along randomly chosen directions. $rp\text{Forests}$ combines the power of ensemble methods [5, 6, 18, 41] and the flexibility of trees. $rp\text{Forests}$ is computationally efficient with a log-linear average complexity for growth and $O(\log(n))$ for search. As the tree partitions the data space

recursively, data points falling in the same tree leaf node would be close to each other or “similar”. This property is leveraged for the construction of the rpf-kernel. Additionally, as individual trees are rpTrees, *rpForests* can adapt to the geometry of the data and readily overcomes the curse of dimensionality [13].

The rpf-kernel is constructed by averaging the incidence matrices induced by trees in *rpForests*. On each tree, an incidence matrix is created with its (i, j) position being 1 if the i^{th} and j^{th} points lie in the same leaf node and 0 otherwise. The rpf-kernel is further transformed by $\exp(S/\beta)$ for some properly chosen bandwidth β to reflect the correct scale at which the data are clustered (just like the Gaussian kernel). *rpfCluster* then works by applying spectral clustering to the rpf-kernel. The cluster membership from spectral clustering is then used, along with a correspondence between the signature point and all points in the same leaf node in DC^2 , to derive the cluster membership for all the original data points. Let T denote the number of trees in *rpForests*. *rpfCluster* is described as Algorithm 3.

Algorithm 3 *rpfCluster*(S)

- 1: Initialize a similarity matrix $K \leftarrow \mathbf{0}$;
 - 2: **for** $i = 1$ to T **do**
 - 3: Grow random projection tree $t_i \leftarrow rpTree(S)$;
 - 4: **for** each leaf node $\mathcal{N} \in t_i$ **do**
 - 5: Increase the similarity count for each entry in $K[\mathcal{N}, \mathcal{N}]$;
 - 6: **end for**
 - 7: **end for**
 - 8: Average K by $K \leftarrow K/T$;
 - 9: $K \leftarrow \exp(K/\beta)$ for some bandwidth β ;
 - 10: Apply spectral clustering to K ;
 - 11: Populate spectral clustering membership to all data points;
-

For details about spectral clustering, the reader can refer to [36, 30, 38].

3 Theoretical analysis when DC^2 is applied to spectral clustering

In applying the DC^2 algorithm, a data point is replaced by a signature point (indeed many points falling into the same leaf node of the rpTrees would be replaced by the same signature point). Due to their discrepancy, there will be an approximation error in the resulting kernel matrix and the subsequent kernel learning. Error bounds can be derived for several kernel learning methods in [11], including kernel ridge regression, support vector machines, graph Laplacian regularization algorithms. We focus specifically here in deriving an error bound when spectral clustering is applied to the kernel learned on the set of signatures generated by DC^2 . The discrepancy is modeled as data perturbation in our analysis.

We treat data perturbation as adding a noise component ϵ to data X [23, 43]

$$\tilde{X} = X + \epsilon. \tag{1}$$

Assume ϵ is symmetric about 0 with bounded support, and let ϵ have standard deviation σ_ϵ that is small compared to σ , the standard deviation for the distribution of X . Note that here we assume that $X \in \mathbb{R}$; this is for simplicity of discussion, and the extension to \mathbb{R}^d is straightforward.

To prepare for the perturbation analysis, let us introduce some notations. Let X_1, \dots, X_N be the given data. Let $\mathcal{K}(\cdot, \cdot)$ be the similarity kernel of interest. Denote $a_{ij} = \mathcal{K}(X_i, X_j)$ and let $A = (a_{i,j})_{i,j=1}^N$ be the similarity matrix. Let \mathcal{L} be the graph Laplacian of A , i.e.,

$$\mathcal{L} = D^{-\frac{1}{2}}(D - A)D^{-\frac{1}{2}}, \quad (2)$$

where $D = \text{diag}(d_1, \dots, d_N)$ with $d_i = \sum_{j=1}^N a_{ij}$, $i = 1, \dots, N$. We will use \sim to denote quantities due to perturbation, e.g., $\tilde{a}_{ij} = \mathcal{K}(\tilde{X}_i, \tilde{X}_j)$. Our analysis is based on an end-to-end error bound w.r.t. the distortion to the Laplacian matrix [23], and a perturbation bound to the Laplacian matrix due to data perturbation [43]. The main result of our perturbation analysis is stated as Theorem 3.1.

Theorem 3.1. *Suppose X_1, \dots, X_N is a sample of bounded support such that $\inf_{1 \leq i \leq N} d_i/N > \delta_0$ holds in probability for some constant $\delta_0 > 0$. Assume the similarity kernel $\mathcal{K}(\cdot, \cdot)$ is uniformly bounded and further there exists universal constant K s.t. $|\mathcal{K}(\tilde{X}_1, \tilde{X}_2) - \mathcal{K}(X_1, X_2)| \leq K \left(\sum_{i=1}^2 \|\tilde{X}_i - X_i\|^2 \right)$ where $\|\cdot\|$ is the Euclidean norm. Assume the data perturbation ϵ is symmetric about 0 with bounded support. Then under suitable technical conditions, the loss ρ in clustering accuracy of a spectral bi-partitioning algorithm due to DC^2 satisfies*

$$\rho \leq \|\tilde{\mathcal{L}} - \mathcal{L}\|_F^2 \leq_p C\sigma_\epsilon^2 + C'\sigma_\epsilon^4,$$

where $\|\cdot\|_F$ indicates the Forbenius norm, for some constants C and C' , as $N \rightarrow \infty$.

Theorem 3.1 states that the loss in clustering accuracy due to data perturbation is bounded by the second and fourth moments of the data approximation error. According to [13], the radius of tree leaves (i.e., the distance between a point and the node centroid) in rpTrees vanishes. Thus the mean data approximation error of the full data by the signature points vanishes and so does the perturbation bound and the loss in clustering accuracy when the number of signature points increases.

4 Related work

There are several lines of work that are related to ours. This includes many work that use the divide-and-conquer principle to tame large scale problems. An influential line of work is *Bag of Little Bootstrap* [25], a big data version of Bootstrap [16]. The idea is to take many very “thin” subsamples to be distributed to many computer nodes, and then results from those individual subsamples are aggregated. [9] considers smoothing spline and explored the tradeoff between computational efficiency and statistical optimality of the *Divide-and-Conquer* methods in a distributed environment. [8] studied penalized regression for data

Data set	# Features	# instances	# classes
Connect-4	42	67,557	3
USCI	37	285,779	2
Cover Type	54	568,772	5
HT Sensor	11	928,991	3
Poker Hand	10	1,000,000	3
Gas Sensor	18	8,386,765	2

Table 1: *A summary of UC Irvine data used in the experiments.*

too big to fit in the memory by working on subsamples of the data and then aggregating the resulting models. Also, [37] learns a quantile function, [47] studies ridge regression, [2] considers the general distributed estimation and inference, [26] learns a Lasso-type linear model at individual sites and then aggregate to de-bias, [31] explores coordinate descent for distributed data, [32] studies the optimality of averaging in distributed computing. Such work are different from ours in that they all *average* results obtained on random subsamples, while we *assemble* results from geometry-aware data partitions with DC^2 algorithm.

Another line of closely related work are those under the term “learning over inherently distributed data” [45, 46]. Instead of dividing the data, these work deal with situations where the data are already distributed, i.e., stored at a number of distributed machines as a result of business operation or diverse data collection channels. Computations are performed on the local data on the machine that stores the data, and then local signatures are sent to a central server for aggregation. Our work almost works in a reverse way, which splits and distributes data to multiple computer cores or machines for parallel computation. Of course, there are also work that use the idea of data compression for approximate large scale computation. This includes fast approximate spectral clustering [43], landmark based spectral clustering [7], and also spectral clustering by the Nyström method [39, 17]. While all use data compression, our method starts with divide-and-conquer and further employs parallel computation.

5 Experiments

Our experiments consists of three parts. We first evaluate the mean squared errors (MSE) in approximating the full data by the collection of signature points when the data is partitioned by random sampling versus by random projection. Then we evaluate $rpfCluster^+$ by comparing it to competing algorithms, including K-means clustering [20] as the baseline, and two algorithms, KASP and RASP, for large scale spectral clustering [43]. In the third part, we evaluate the performance of $rpfCluster^+$ under different data partition schemes and different number of partitions. We start by describing the data, the performance metrics and competing methods.

We use 6 benchmark datasets from the UC Irvine Machine Learning Repository [27], including the Connect-4, USCI (US Census Income), Cover type, HT

Sensor, Poker Hand, and the Gas Sensor data. Table 1 is a summary of the datasets. For Connect-4, USCI, and Poker Hand data, we follow procedures described in [43] to preprocess the data. The original *USCI* data has 299,285 instances with 41 features. We exclude features #26, #27, #28 and #30, due to too many missing values, and then remove all instances with missing values. This leaves 285,799 instances on 37 features, with all categorical variables converted to integers. The original *Cover Type* data has 581,012 instances. We excluded the two small classes (i.e., 4 and 5) for fast evaluation of accuracy (otherwise all 7! permutations need to be evaluated, but that is not the focus of the present work), and this leaves 568,772 instances; we also standardized each of the first 10 features to have a mean 0 and variance 1. The original *Poker Hand* data is highly unbalanced, with 6 small classes containing less than 1% of the data. Merging small classes gives 3 final classes with a class distribution of 50.12%, 42.25% and 7.63%, respectively. The *Gas Sensor* data consists of two different gas mixtures: Ethylene mixed with CO, and Ethylene mixed with Methane. The two different gas mixtures form two classes in the data. The Connect-4, the USCI, and the Gas Sensor data are standardized on all features.

The performance is assessed by clustering accuracy and the computation time. The use of clustering accuracy aligns closely to the ultimate goal of clustering—assigning data points to proper groups. In contrast, many other performance metrics are often a surrogate of this due to the lack of true labels. However, as we are evaluating clustering algorithms, we have the freedom to use data with true labels.

Definition. Let $\mathcal{L} = \{1, 2, \dots, l\}$ be the label set. Let $h(\cdot)$ and $\hat{h}(\cdot)$ be the true label and the label obtained by a clustering algorithm, respectively. The *clustering accuracy* is defined as

$$\rho_c(\hat{h}) = \max_{\tau \in \Pi_{\mathcal{L}}} \left\{ \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{\tau(h(X_i)) = \hat{h}(X_i)\} \right\}, \quad (3)$$

where \mathbb{I} is the indicator function and $\Pi_{\mathcal{L}}$ is the set of all permutations on the label set \mathcal{L} . It measures the fraction of labels by a clustering algorithm that agree with the true labels that come with the dataset up to a permutation of the true labels. This is a natural extension of the classification accuracy (under 0-1 loss) and has been used by many work in clustering [40, 29, 43].

5.1 Competing methods and parameters

We compare *rpfCluster*⁺ to three other clustering algorithms—K-means clustering, and two variants of fast approximate spectral clustering algorithms, KASP and RASP [43]. *Note that here our goal is not to show which algorithm outperforms others, but to demonstrate that our DC² algorithm can significantly speed up large scale kernel learning (with a data-driven similarity kernel).*

K-means clustering [28] is one of the most widely used clustering algorithms. It starts with randomly generated cluster centroids, and then alternates between two steps: 1) assign data points to the closest centroid; and 2) recalculate the cluster centroid for each cluster, until the change to the within-cluster sum of

squares is small enough. The R package *kmeans()* is used with the “Hartigan-Wong” [20] initialization, and the maximum number of iterations and the number of restarts are set to be (200, 20).

KASP and RASP are among the fastest algorithms for spectral clustering. Both are based on the idea of shifting expensive spectral clustering to a small amount of structure-preserving data signatures. While KASP obtains data signatures via K-means clustering, RASP grows rpTrees. The data compression ratio is chosen such that the signature set has about 500-1000 points. The bandwidth parameter for the Gaussian kernel varies with a step size of 0.1 in the range [0.1, 1] and 1 in (1, 200].

For DC^2 , the data compression ratio is such that the collection of signature points has a size about 1000. For *rpfCluster*, the number of trees is fixed at 800, the node splitting constant parameter n_s is fixed at 30, and the step size for the search of bandwidth β is 1 in the range [10, 80]. All results in our experiments are averaged over 100 runs.

5.2 Partition by random sampling or projection

We advocate the use of recursive random projections for data partitions in the DC^2 algorithm. In Section 2.1, we argue that this would lead to more robust algorithms for large scale kernel learning, and show that the loss in accuracy due to data reduction in DC^2 is upper bounded by the approximation error of the full data by the collection of signature points. In this section, we will empirically demonstrate that partition by random projections tends to have a smaller mean squared approximation error. When the difference becomes significant (e.g., the ratio of MSEs is larger than 1.5), it will translate to a noticeable difference in clustering accuracy (c.f. Section 5.4).

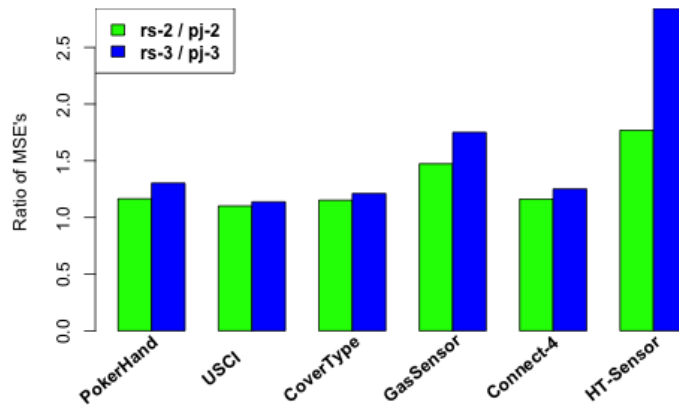


Figure 2: *Ratio of MSEs in DC^2 when dividing the data by random sampling (indicated by “rs”) and random projection (indicated by “pj”). The number of data partitions are indicated in the legend.*

Figure 2 shows the ratio of MSEs due to the DC^2 algorithm when dividing the data by random sampling and recursive random projections. It can be seen that, in all cases, the MSE by random sampling is larger than that by random projections. For two data sets, HT sensor and the Gas sensor, the ratios are larger than 1.5. Later in Section 5.4, we will see that random sampling leads to a remarkably more loss in clustering accuracy on these two datasets. This gives support to our choice of recursive random projections for data partitions in the DC^2 algorithm.

5.3 Comparison of different clustering algorithms

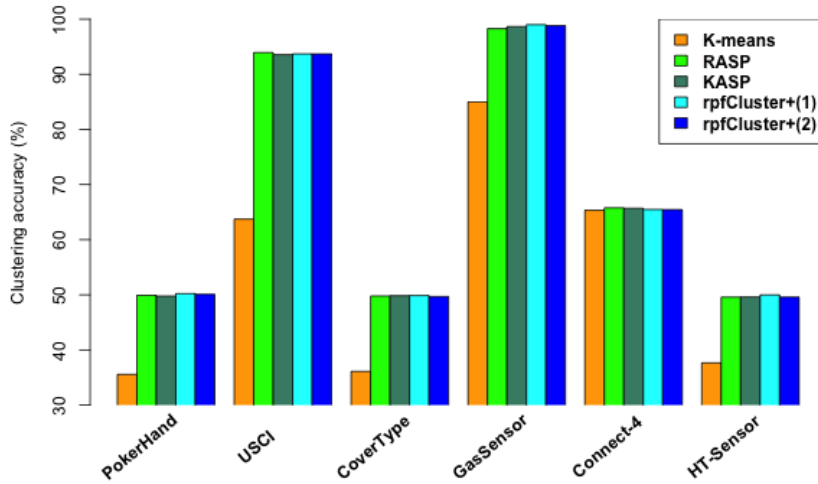


Figure 3: Clustering accuracy under K -means clustering, RASP, KASP, and $rpfCluster^+$ (the numbers in the parenthesis indicate the number of partitions, and the partitions are obtained by random projections).

We compare the performance of $rpfCluster^+$ to three competing algorithms, K -means clustering, KASP and RASP, on both clustering accuracy and computation time. These are shown in Figure 3 and Figure 4, respectively. It can be seen that the clustering accuracy of K -means clustering is much lower than the other three algorithms on all but one dataset, while that of the other three are quite similar. For computation time, K -means clustering is the fastest on all the datasets while KASP is the slowest on almost all the data. The computation time by $rpfCluster^+$ (with one data partition) is close to that of RASP on all the data, and by running over two data partitions, $rpfCluster^+$ becomes faster than RASP, and is close to K -means clustering on most of the datasets.

5.4 Performance under varying number of partitions

We also evaluate the performance of $rpfCluster^+$ under different schemes of data partitions, random sampling and random projection, with 2 or 3 partitions.

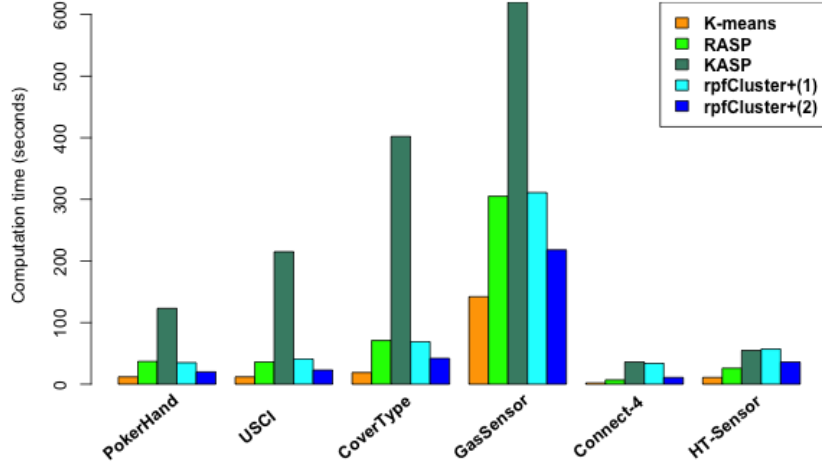


Figure 4: Clustering time under K -means clustering, RASP, KASP, and $rpfCluster^+$ (the numbers in the parenthesis indicate the number of partitions, and the partitions are obtained by random projections).

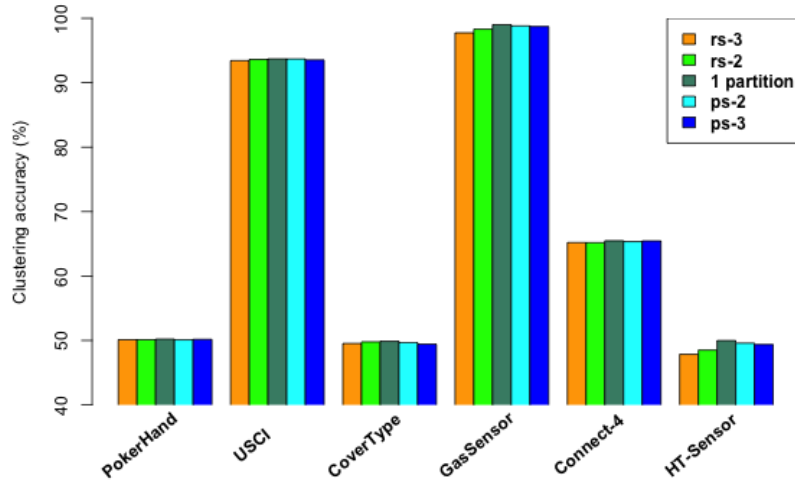


Figure 5: Clustering accuracy under different schemes of dividing the data and with different partitions for $rpfCluster^+$. The number in the legend indicates the number of partitions.

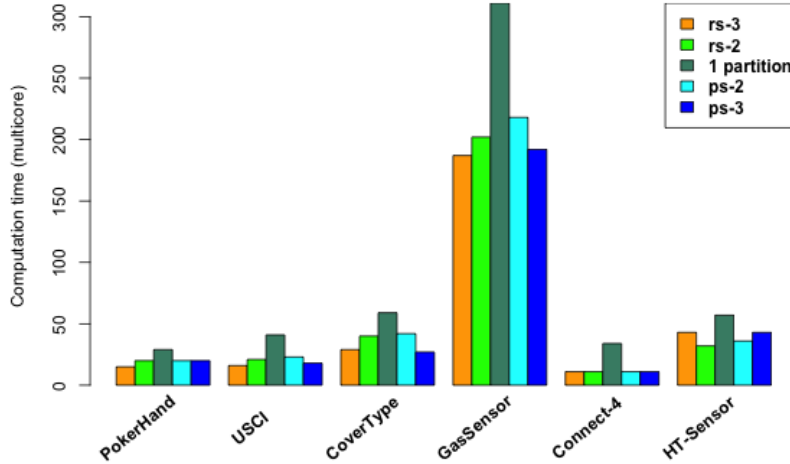


Figure 6: *Clustering time under different schemes of dividing the data and with different partitions for $rpfCluster^+$ running in multicore mode. The number in the legend indicates the number of partitions.*

Figure 5 and Figure 6 show the clustering accuracy and computation time, respectively. In general, more data partitions leads to a shorter computation time; the gain in computational efficiency starts diminishing when more partitions are applied on smaller data. It is worth noting is that, on two datasets, the Gas sensor data and the HT-sensor data, the decrease in accuracy with more data partitions becomes noticeable when data partitions are obtained by random sampling while the loss of accuracy by random projections remains negligible. We attribute this to the much higher MSEs by random sampling, as discussed in Section 5.2.

An additional experiment is conducted when $rpfCluster^+$ is running in sequential mode, that is, assume there is only one core or a single machine in the cluster. Figure 7 shows that there is a potential advantage in computational efficiency to apply the DC^2 algorithm to large scale problems, even if there is no parallel infrastructure (due possibly to the non-linearity of the underlying kernel learning algorithm).

6 Conclusions

We have proposed an effective algorithm, DC^2 , for large scale kernel learning. DC^2 applies divide-and-conquer with a further distortion minimizing representation compression on the data, and achieves the efficiency of the conventional sampling based approach for large scale computation with a potential of parallel computation on multicore or clustered computers. With DC^2 , the partition and the subsequent representation compression of data are implemented under a unified operation—recursive random projections. We advocate the use of re-

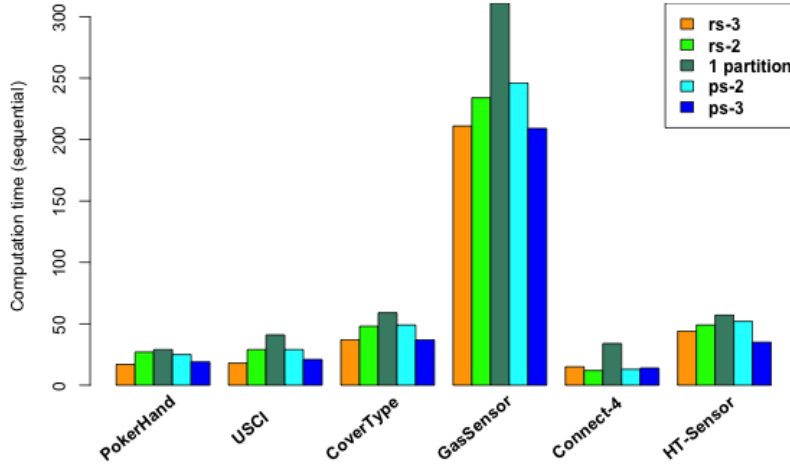


Figure 7: Clustering time under different schemes of dividing the data and with different partitions for $rpfCluster^+$ running in sequential mode. The number in the legend indicates the number of partitions.

cursive random projections for dividing the data in divide-and-conquer, which has the advantage of smaller MSEs compared to the conventional approach of random sampling. On a random projection forests based clustering algorithm, we demonstrated the power and efficiency of DC^2 algorithm (resulting algorithm termed as $rpfCluster^+$). $rpfCluster^+$ achieves a similar level of clustering accuracy as KASP and RASP, some of the fastest approximate spectral clustering algorithms, and has a running time close to that of K-means clustering. Theoretical analysis is carried out on DC^2 when the resulting data signatures are used as input to spectral clustering, and we show that the loss in accuracy due to data reduction is upper bounded by the data approximation error which would vanish with recursive random projections. Due to the easy implementation and flexibility of DC^2 , we expect it to be applicable to general large scale learning and inference problems.

References

- [1] F. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2003.
- [2] H. Battay, J. Fan, H. Liu, J. Lu, and Z. Zhu. Distributed estimation and inference with statistical guarantees. *arXiv:1509.05457*, 2015.
- [3] M. Belkin, I. Matveeva, and P. Niyogi. Regularization and semi-supervised learning on large graphs. In *International Conference on Computational Learning Theory (COLT)*, 2004.
- [4] J. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

- [5] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.
- [6] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.
- [7] X. Chen and D. Cai. Large scale spectral clustering with landmark-based representation. In *AAAI*, 2011.
- [8] X. Chen and M. Xie. A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, 24:1655–1684, 2014.
- [9] G. Cheng and Z. Shang. Computational limits of divide-and-conquer method. *arXiv:1512.09226*, 2015.
- [10] J. W. Cooley and J. W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301, 1965.
- [11] C. Cortes, M. Mohri, and A. Talwalkar. On the impact of kernel approximation on learning accuracy. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 113–120, 2010.
- [12] C. Cortes and V. N. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.
- [13] S. Dasgupta and Y. Freund. Random projection trees and low dimensional manifolds. In *Fortieth ACM Symposium on Theory of Computing (STOC)*, 2008.
- [14] I. Dhillon, Y. Guan, and B. Kulis. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM international conference on Knowledge discovery and data mining (SIGKDD)*, 2004.
- [15] P. Drineas and M. W. Mahoney. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. In *Proceedings of COLT*, pages 323–337, 2005.
- [16] B. Efron. Bootstrap methods: another look at the Jackknife. *Annals of Statistics*, 7(1):1–28, 1979.
- [17] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the Nyström method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):214–225, 2004.
- [18] Y. Freund and R. Schapire. Experiments with a new boosting algorithm. In *International Conference on Machine Learning (ICML)*, 1996.
- [19] K. Fukumizu, F. Bach, and A. Gretton. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8:361–383, 2007.
- [20] J. A. Hartigan and M. A. Wong. A K-means clustering algorithm. *Applied Statistics*, 28(1):100–108, 1979.
- [21] C. A. R. Hoare. Quicksort. *The Computer Journal*, 5(1):10–16, 1962.
- [22] T. Hofmann, B. Schölkopf, and A. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220, 2008.

- [23] L. Huang, D. Yan, M. I. Jordan, and N. Taft. Spectral clustering with perturbed data. In *Advances in Neural Information Processing Systems (NIPS)*, volume 21, pages 705–712, 2009.
- [24] A. Karatsuba and Yu. Ofman. Multiplication of many-digital numbers by automatic computers. *Proceedings of the USSR Academy of Sciences*, 145:293–294, 1962.
- [25] A. Kleiner, A. Talwalkar, P. Sarkar, and M. I. Jordan. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816, 2014.
- [26] J. D. Lee, Q. Liu, Y. Sun, and J. E. Taylor. Communication-efficient sparse regression. *Journal of Machine Learning Research*, 18:1–30, 2017.
- [27] M. Lichman. UC Irvine Machine Learning Repository. <http://archive.ics.uci.edu/ml>, 2013.
- [28] S. P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(1):128–137, 1982.
- [29] M. Meila, S. Shortreed, and L. Xu. Regularized spectral learning. Technical report, Department of Statistics, University of Washington, 2005.
- [30] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: analysis and an algorithm. In *Neural Information Processing Systems (NIPS)*, volume 14, 2002.
- [31] P. Richtarik and M. Takac. Distributed coordinate descent method for learning with big data. *Journal of Machine Learning Research*, 17:1–25, 2016.
- [32] J. D. Rosenblatt and B. Nadler. On the optimality of averaging in distributed statistical learning. *Information and Inference: A Journal of the IMA*, 5(4):379–404, 2016.
- [33] C. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML)*, 1998.
- [34] B. Schölkopf. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.
- [35] B. Schölkopf and A. Smola. *Learning with kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001.
- [36] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
- [37] S. Volgushev, S.-K. Chao, and G. Cheng. Distributed inference for quantile regression processes. *The Annals of Statistics*, 47(3):1634–1662, 2019.
- [38] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, pages 395–416, 2007.

- [39] C. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, 2001.
- [40] E. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *Proceedings of Neural Information Processing Systems (NIPS)*, pages 521–528, 2002.
- [41] D. Yan, A. Chen, and M. I. Jordan. Cluster Forests. *Computational Statistics and Data Analysis*, 66:178–192, 2013.
- [42] D. Yan, S. Gu, Y. Xu, and Z. Qin. Similarity kernel and clustering via random projection forests. *arXiv:1908.10506*, 2019.
- [43] D. Yan, L. Huang, and M. I. Jordan. Fast approximate spectral clustering. *Technical Report 772, Department of Statistics, UC Berkeley*, 2009.
- [44] D. Yan, Y. Wang, J. Wang, H. Wang, and Z. Li. K-nearest neighbor search by random projection forests. *arXiv:1812.11689*, 2018.
- [45] D. Yan, Y. Wang, J. Wang, G. Wu, and H. Wang. Fast communication-efficient spectral clustering over distributed data. *arXiv:1905.01596*, 2019.
- [46] D. Yan and Y. Xu. Learning over inherently distributed data. *arXiv:1907.13208*, 2019.
- [47] Y. Zhang, J. Duchi, and M. Wainwright. Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16(1):3299–3340, 2015.