

# k-averages, a linear optimization alternative to kernel k-means

Mathias Rossignol      Mathieu Lagrange

November 19, 2014

## Abstract

We present an iterative flat clustering algorithm designed to operate on arbitrary similarity matrices, provided, with the only constraint that these matrices be symmetrical. Although functionally very close to kernel k-means, our proposal performs an maximization of average intra-class similarity, instead of a squared distance minimization, in order to remain closer to the semantics of similarities. We show that this approach allows relaxing the conditions on usable matrices, as well as opening better optimization possibilities. Systematic evaluation on a variety of data sets shows that the proposed approach equals or outperforms kernel k-means in most cases.

## 1 Introduction

## 2 Related Works

Kernel k-means (cite Girolami) has been for the past few years an algorithm of choice for flat data clustering with known number of clusters (cite salient uses of kmeans). It makes use of a mathematical technique known as the “kernel trick” to extend the classical k-means clustering algorithm (cite original kmeans paper) to criteria beyond simple euclidean distance proximity. Since it constitutes the closest point of comparison with our own work, we will present it here in detail, before moving on to the description of our algorithm, before dedicating a section to the comparison of the two approaches.

In the case of kernel k-means, the kernel trick consists in considering that the k-means algorithm is operating in an unspecified, possibly very high-dimensional Euclidian space; but instead of specifying the properties of that space, and the coordinates of objects in it, the equations governing the algorithm are modified so that everything can be computed knowing only the scalar products between points. The symmetrical matrix containing those scalar products is known as a kernel, noted  $\mathcal{K}$ .

### 2.1 Kernel k-means objective function

In this section and the following, we shall adopt the following convention:  $N$  is the number of objects to cluster and  $C$  the number of clusters;  $N_c$  is the

number of objects in cluster  $c$ , and  $\mu_c$  is the centroid of that cluster.  $z_{cn}$  is the membership function, whose value is 1 if object  $o_n$  is in class  $c$ , 0 otherwise.

Starting from the objective function minimized by the k-means algorithm, expressing the sum of squared distances of points to the centroids of their respective clusters:

$$S = \sum_{c=1}^C \sum_{n=1}^N z_{cn} (o_n - \mu_c) (o_n - \mu_c)^\top$$

And using the definition of centroids as:

$$\mu_c = \frac{1}{N_c} \sum_{n=1}^N z_{cn} o_n$$

$S$  can be developed and rewritten in a way that does not explicitly refer to the centroid positions, since those cannot be computed:

$$S = \sum_{c=1}^C \sum_{n=1}^N z_{cn} Y_{cn}$$

where

$$Y_{cn} = (o_n - \mu_c) (o_n - \mu_c)^\top \quad (1)$$

$$= o_n \cdot o_n - 2o_n \cdot \mu_c + \mu_c \cdot \mu_c \quad (2)$$

$$= o_n \cdot o_n - 2o_n \cdot \frac{1}{N_c} \sum_{i=1}^N z_{ci} o_i + \left( \frac{1}{N_c} \sum_{i=1}^N z_{ci} o_i \right) \cdot \left( \frac{1}{N_c} \sum_{i=1}^N z_{ci} o_i \right) \quad (3)$$

$$= o_n \cdot o_n - \frac{2}{N_c} \sum_{i=1}^N z_{ci} o_n \cdot o_i + \frac{1}{N_c^2} \sum_{i=1}^N \sum_{j=1}^N z_{ki} z_{kj} o_i \cdot o_j \quad (4)$$

$$= \mathcal{K}_{nn} - \frac{2}{N_c} \sum_{i=1}^N z_{ci} \mathcal{K}_{ni} + \frac{1}{N_c^2} \sum_{i=1}^N \sum_{j=1}^N z_{ki} z_{kj} \mathcal{K}_{ij} \quad (5)$$

Since the sum of  $\mathcal{K}_{nn}$  over all points remains constant, and the sum of squared centroid norms (third, quadratic, term of Equation 5) is mostly bounded by the general geometry of the cloud of objects, we can see that minimizing this value implies maximizing the sum of the central terms, which are the average scalar products of points with other points belonging to the same class. Given a similarity matrix possessing the necessary properties to be considered as a kernel matrix (positive semidefinite), the kernel k-means algorithm can therefore be used to create clusters that somehow maximize the average intra-cluster similarity.

## 2.2 Algorithm

Finding the configuration that absolutely minimizes  $S$  (eq 2.1) is an NP-hard problem; however several approaches allow finding a good approximate result. We shall only focus here on the fastest and most popular, an iterative assignment / update procedure commonly referred to as the "k-means algorithm", or as a discrete version of Lloyd's algorithm, detailed in Algorithm 1.

**Data:** number of objects  $N$ , number of classes  $C$ , kernel matrix  $\mathcal{K}$   
**Result:** label vector  $L$  defining a partition of the objects into  $C$  classes

```

1 Initialization: fill  $L$  with random values in  $[1..C]$ ;
2 while  $L$  is modified do
3   for  $n \leftarrow 1$  to  $N$  do
4     for  $c \leftarrow 1$  to  $C$  do
5       | Compute  $Y_{cn}$  following eq 5 (note:  $z_{cn} = (L_n == c) ? 1 : 0$ )
6     end
7      $L_n = \operatorname{argmin}_c(Y_{cn})$ ;
8   end
9 end

```

**Algorithm 1:** Lloyd’s algorithm applied to minimizing the kernel k-means objective.

### 3 Foundations of the k-averages algorithm

Our proposal with the k-averages algorithm is to adopt an alternative objective function that does not rely on a geometric interpretation, like kernel k-means, but on a more simple, direct understanding of a similarity matrix. The goal is to maximize the average intra-cluster similarity between points, a commonly used metric to evaluate clustering quality, and one whose computation is very simple—linear in time.

Due to its simplicity, however, that objective function cannot simply be “plugged into” the standard kernel k-means algorithm: it lacks the geometric requisites to ensure convergence. We must therefore propose an adapted algorithmic framework to exploit it: first, we show here that it is possible to easily compute the impact on the global objective function of moving a single point from one class to another; we then introduce an algorithm intended to take advantage of that formula.

#### 3.1 Conventions and possible objective functions

In addition to the notations already presented above, we index here the set of elements belonging to a given cluster  $c_k$  as  $c_k = \{o_{k1}, \dots, o_{kN_k}\}$ . To simplify below, when we’re simply considering one class, no matter which, we shall omit the first index and write  $c = \{o_1, \dots, o_{N_c}\}$ .

The similarity between objects shall be written  $s(o_i, o_j)$ . Let us extend the notation  $s$  to the *similarity of an object to a class*, which we define as the average similarity of that object with all objects of the class.  $s(o, c)$  accepts two definitions, depending on whether or not  $o$  is in  $c$ :

If  $o \notin c$ ,

$$s(o, c) = \frac{1}{N_c} \sum_{i=1}^{n_c} s(o, o_i) \quad (6)$$

If  $o \in c$ , then necessarily  $\exists i \mid o = o_i$

$$s(o, c) = s(o_i, c) = \frac{1}{N_c - 1} \sum_{j=1 \dots n_c, j \neq i} s(o_i, o_j) \quad (7)$$

Let's call "quality" of a class the average intra-class object-to-object similarity, and write it  $\mathcal{Q}$ :

$$\mathcal{Q}(c) = \frac{1}{N_c} \sum_{i=1}^{n_c} s(o_i, c) \quad (8)$$

We do not, in our work, explicitly refer to class centroids, preferring to directly consider averages of similarity values between individuals within clusters. Indeed, we never refer to a geometrical interpretation of the data. However, it should be noted that since in k-means (and kernel k-means) the centroid of a class is defined as an average of all points in that class,  $\mathcal{Q}$  is strictly equivalent to the average point to centroid similarity.

Using the notations above, we define our objective function as the average class quality, either normalized by class:

$$O_1 = \frac{1}{C} \sum_{i=1}^C \mathcal{Q}(c_i)$$

Or normalized by object:

$$O_2 = \frac{1}{N} \sum_{i=1}^C N_i \mathcal{Q}(c_i)$$

We will now compute, for both of those values, the impact on the global objective function of moving one object from one class to another. We will then show how, by using those formulas as a guide for the optimal reallocation of objects, and only moving objects that have a strictly positive impact on the function, we can guarantee the convergence of an iterative algorithm attempting to maximize those objective functions.

### 3.2 Naive k-averages

### 3.3 Impact of object reallocation on class quality

Considering a class  $c$ , let us develop the expression of  $\mathcal{Q}(c)$  into a more useful form. Since all objects are in  $c$ , we use the formula in (7) to get:

$$\begin{aligned} \mathcal{Q}(c) &= \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{1}{N_c - 1} \sum_{\substack{j=1 \dots N_c \\ j \neq i}} s(o_i, o_j) \\ &= \frac{1}{N_c(N_c - 1)} \sum_{i=1}^{N_c} \sum_{\substack{j=1 \dots N_c \\ j \neq i}} s(o_i, o_j) \end{aligned} \quad (9)$$

Using the assumption that the similarity matrix is symmetrical, we can reach (this is an indispensable transformation for future calculations):

$$\mathcal{Q}(c) = \frac{2}{N_c(N_c - 1)} \sum_{i=2}^{N_c} \sum_{j=1}^{i-1} s(o_i, o_j) \quad (10)$$

For future use, let us define the notation:

$$\Sigma(c) = \sum_{i=2}^{N_c} \sum_{j=1}^{i-1} s(o_i, o_j) \quad (11)$$

Thus:

$$\mathcal{Q}(c) = \frac{2}{N_c(N_c - 1)} \Sigma(c) \quad \text{and} \quad \Sigma(c) = \frac{N_c(N_c - 1)\mathcal{Q}(c)}{2} \quad (12)$$

### 3.3.1 Removing an object from a class

Assuming that  $o \in c$ , necessarily  $\exists i \mid o = o_i$ . Since the numbering of objects is arbitrary, we can assume that  $o = o_{N_c}$  then generalize from the result thus obtained.

$$\begin{aligned} \mathcal{Q}(c \setminus o_{N_c}) &= \frac{2}{(N_c - 1)(N_c - 2)} \sum_{i=2}^{N_c-1} \sum_{j=1}^{i-1} s(o_i, o_j) \\ &= \frac{2}{(N_c - 1)(N_c - 2)} \left[ \Sigma(c) - \sum_{j=1}^{N_c-1} s(o_{N_c}, o_j) \right] \\ &= \frac{2}{(N_c - 1)(N_c - 2)} [\Sigma(c) - (N_c - 1)s(o_{N_c}, c)] \\ &= \frac{2N_c(N_c - 1)\mathcal{Q}(c)}{2(N_c - 1)(N_c - 2)} - \frac{2(N_c - 1)s(o_{N_c}, c)}{(N_c - 1)(N_c - 2)} \\ &= \frac{N_c\mathcal{Q}(c) - 2s(o_{N_c}, c)}{N_c - 2} \end{aligned} \quad (13)$$

The quality of a class after removal of an object is thus:

$$\mathcal{Q}(c \setminus o) = \frac{N_c\mathcal{Q}(c) - 2s(o, c)}{N_c - 2} \quad (14)$$

And the change in quality from its previous value:

$$\begin{aligned} \mathcal{Q}(c \setminus o) - \mathcal{Q}(c) &= \frac{N_c\mathcal{Q}(c) - (N_c - 2)\mathcal{Q}(c) - 2s(o, c)}{N_c - 2} \\ &= \frac{2(\mathcal{Q}(c) - s(o, c))}{N_c - 2} \end{aligned} \quad (15)$$

### 3.3.2 Adding an object to a class

Assuming that  $o \notin c$ , we can similarly to what has been done previously (numbering is arbitrary) consider for the sake of simplicity that  $o$  becomes  $o_{N_c+1}$  in the modified class  $c$ . Following a path similar to above, we get:

$$\begin{aligned}
\mathcal{Q}(c \cup o_{N_c+1}) &= \frac{2}{N_c(N_c+1)} \sum_{i=2}^{N_c+1} \sum_{j=1}^{i-1} s(o_i, o_j) \\
&= \frac{2}{N_c(N_c+1)} [\Sigma(c) + N_c s(o_{N_c+1}, c)] \\
&= \frac{(N_c-1)\mathcal{Q}(c) + 2s(o_{N_c+1}, c)}{N_c+1}
\end{aligned} \tag{16}$$

The quality of a class  $c$  after adding an object  $o$  is thus:

$$\mathcal{Q}(c \cup o) = \frac{(N_c-1)\mathcal{Q}(c) + 2s(o, c)}{N_c+1} \tag{17}$$

And the change in quality from its previous value:

$$\begin{aligned}
\mathcal{Q}(c \cup o) - \mathcal{Q}(c) &= \frac{(N_c-1)\mathcal{Q}(c) - (N_c+1)\mathcal{Q}(c) + 2s(o, c)}{N_c+1} \\
&= \frac{2(s(o, c) - \mathcal{Q}(c))}{N_c+1}
\end{aligned} \tag{18}$$

### 3.4 Impact of object reallocation on the global objective function

The influence these changes in class quality have on the global objective function depends upon what normalization we choose to adopt.

#### 3.4.1 Class-normalized objective function

In that case, the calculation is direct: from (15) and (18), we can see that the impact on the objective function of moving an object  $c$  from class  $c_s$  (“source”), to whom it belongs, to a distinct class  $c_t$  (“target”) is:

$$\delta_o(c_s, c_t) = \frac{2(s(o, c_t) - \mathcal{Q}(c_t))}{N_t+1} + \frac{2(\mathcal{Q}(c_s) - s(o, c_s))}{N_s-2} \tag{19}$$

#### 3.4.2 Object-normalized objective function

This complicates the calculation a bit, but not much: when moving an object  $o$  from class  $c_s$  (“source”), to whom it belongs, to a distinct class  $c_t$  (“target”),  $(N_s-1)$  objects are affected by the variation in (15), and  $N_t$  are affected by that in (18), in addition to the variation in similarity of  $o$  to the class it belongs to:

$$\delta_o(c_s, c_t) = \frac{2N_t(s(o, c_t) - \mathcal{Q}(c_t))}{N_t+1} + \frac{2(N_s-1)(\mathcal{Q}(c_s) - s(o, c_s))}{N_s-2} + s(o, c_t) - s(o, c_s) \tag{20}$$

In both cases, computing that impact is a fixed-cost operation; we can therefore use those formulas as the basis for an efficient iterative algorithm.

## 4 K-averages algorithm

Our approach does not allow us to benefit, like kernel k-means, from the convergence guarantee brought by the geometric foundation of k-means. In consequence, we cannot apply a "batch" approach where at each iteration all elements are moved to their new class, and all distances (or similarities) are computed at once. Therefore, for each considered object, after finding its ideal new class, we must update the class properties for the two modified classes (source and destination), as well as recompute the average class-object similarities for them.

Although this seems at first like systematically updating everything at each object re-allocation should have a huge performance impact, our reliance on simple averages without any quadratic terms makes it possible to have very simple update formulas: new class qualities are given by Equations 14 and 17, and new object-class similarities can be computed by:

$$\begin{aligned} s(i, c_s(t+1)) &= \frac{N_s(t) \cdot s(i, c_s(t)) + s(i, n)}{N_s(t) + 1} \\ s(i, c_t(t+1)) &= \frac{N_t(t) \cdot s(i, c_t(t)) - s(i, n)}{N_t(t) - 1} \end{aligned} \quad (21)$$

where  $i$  is any object index,  $n$  is the recently reallocated object,  $c_s$  the "source" class that object  $i$  was removed from, and  $c_t$  the "target" class that object  $n$  was added to.

The full description of k-averages is given in Algorithm 2.

**Data:** number of objects  $N$ , number of classes  $C$ , similarity matrix  $\mathcal{S}$

**Result:** label vector  $L$  defining a partition of the objects into  $C$  classes

```

1 Initialization: Fill  $L$  with random values in  $[1..C]$ ;
2 Compute initial class qualities  $\mathcal{Q}$  following eq 10;
3 Compute initial object-class similarities  $\mathcal{S}$  following eq 7 or eq 6;
4 while  $L$  is modified do
5   for  $i \leftarrow 1$  to  $N$  do
6     previousClass  $\leftarrow L_i$ ;
7     nextClass  $\leftarrow \operatorname{argmin}_k \delta_i(\text{previousClass}, k)$  (following the definition
      of  $\delta$  in eq 19 or 20);
8     if nextClass  $\neq$  previousClass then
9        $L_i \leftarrow$  nextClass;
10      Update  $\mathcal{Q}_{\text{previousClass}}$  following eq 14;
11      Update  $\mathcal{Q}_{\text{nextClass}}$  following eq 17;
12      for  $j \leftarrow 1$  to  $N$  do
13        Update  $\mathcal{S}(j, \text{nextClass})$  and  $\mathcal{S}(j, \text{previousClass})$ 
14        following eq 21;
15      end
16    end
17  end
18 end

```

**Algorithm 2:** K-averages algorithm.

## 5 Complexity analysis

We study in this section the complexity of the two approaches presented above, first from the point of view of raw complexity, then focusing on memory access.

### 5.1 Computational complexity

#### 5.1.1 Kernel k-means

As can be seen on Algorithm 1, the operation on line 5 is the most costly part of the algorithm: for each object  $n$  and class  $c$ , at each iteration, it is necessary to compute  $Y_{cn}$  from Equation 5—an  $O(N^2)$  operation in itself, per object. The impossibility of simply computing the distances to a known centroid as is done in simple k-means, gives kernel k-means a much higher complexity, globally  $O(N^3)$  per iteration, independently of how many objects are moved for that iteration.

It is, however, possible to improve the performance of kernel k-means by noting that in Equation 5, the third term of the equation, which has the highest complexity, is only dependent on class definitions, and not on the considered object. We can therefore rewrite Equation 5 as:

$$Y_{cn} = \mathcal{K}_{nn} - \frac{2}{N_c} \sum_{i=1}^N z_{ci} \mathcal{K}_{ni} + M_c \quad (22)$$

where

$$M_c = \frac{1}{N_c^2} \sum_{i=1}^N \sum_{j=1}^N z_{ci} z_{cj} \mathcal{K}_{ij} \quad (23)$$

Algorithm 1 thus becomes Algorithm 3, where the values of  $M_c$  are computed once at the beginning of each loop (line 4) then reused on line 8, thus reducing the overall complexity to  $O(n^2)$  per iteration.

#### 5.1.2 K-averages

For the k-averages method presented as Algorithm 2, the complexity of each iteration is

- $O(NC')$  corresponding to the best class search at line 7
- $O(NM)$  corresponding to the object-to-class similarity update at line 13, where  $M$  is the number of objects moved at a given iteration.

In the worst case scenario,  $M = N$ , and the complexity for one iteration of the algorithm remains the same as for the optimized kernel k-means algorithm,  $O(n^2)$ . In practice, however, as can be seen on Figure 5.1.2, the number of objects moving from one class to another decreases sharply after the first iteration, meaning that the complexity of one iteration becomes quickly much lower than  $O(n^2)$ . Thus, while the first iteration of k-averages has a similar complexity with kernel k-means, the overall cost of a typical run of the algorithm (from 10 to 50 iterations) is much lower.



**Data:** number of objects  $N$ , number of classes  $C$ , kernel matrix  $\mathcal{K}$

**Result:** label vector  $L$  defining a partition of the objects into  $C$  classes

```

1 Initialization: fill  $L$  with random values in  $[1..C]$ ;
2 while  $L$  is modified do
3   for  $c \leftarrow 1$  to  $C$  do
4     | Compute  $M_c$  following eq 23
5   end
6   for  $n \leftarrow 1$  to  $N$  do
7     | for  $c \leftarrow 1$  to  $C$  do
8       | Compute  $Y_{cn}$  following eq 22 (note:  $z_{cn} = (L_n == c) ? 1 : 0$ )
9     | end
10    |  $L_n = \operatorname{argmin}_c(Y_{cn})$ ;
11  end
12 end

```

**Algorithm 3:** Lloyd's algorithm applied to minimizing the kernel k-means objective, optimized version.

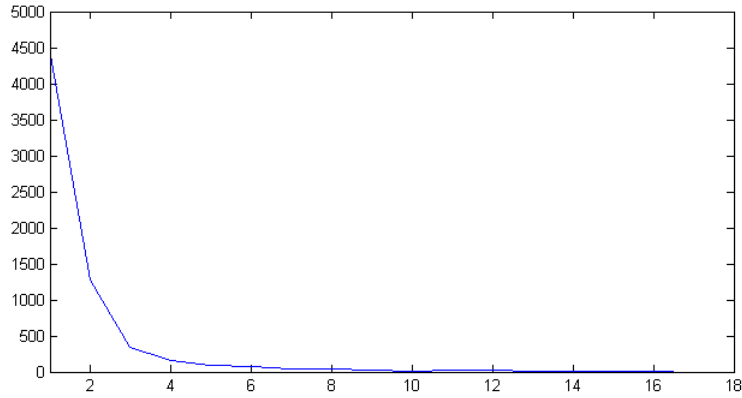


Figure 1: Number of moved objects per iteration during a typical clustering of 5000 objects.

## 5.2 Memory access

The lowered computational costs is also accompanied by a diminution in memory access: as can be seen from Equation 21, in order to compute the new object-to-class similarities after moving an object  $n$ , only line  $n$  of the similarity matrix needs to be read. For the rest of the algorithm, only the (much smaller) object-to-class similarity matrix is used. By contrast, in the case of kernel k-means, the computation of  $M_c$  values at each iteration require that the whole similarity matrix be read, which can be a serious performance bottleneck in the case of large object collections.

Moreover, the similarity update function of k-averages, by reading one line of the matrix at a time, presents good data locality properties, which make it play well with standard memory paging strategies.

## 5.3 Actual running time

To further study the difference in execution speed between the two approaches, we have written simple C implementations of Algorithms 3 and 2, with minimal operational overhead: both read the similarity matrix from a binary file where all matrix values are stored sequentially in standard reading order, line by line, and write out the result of the clustering as a label text file.

## 6 Comparison of approaches

(copy-pasted from the end of the former kmeans section)

However, it should be noted that this is an indirect result of a squared distance minimization objective, and that the impact of the third term of Eq 5 on the the minimized value is hard to quantify. This raises question as to the actual semantics of algorithms based on that objective function, and concerning the exact connection between the initial intent (similarity-based clustering) an the obtained result.

Moreover, if a Euclidean geometric interpretation of the studied objects is not trivial—which is likely to be the case, otherwise the use of a simple k-means could be considered—then similarity values gathered into a matrix may not necessarily make a positive semi definite matrix, and thus not a proper kernel. To solve that problem, (Dhillon-Kulis) suggest, following (Roth-Laub), to offset all diagonal elements of the matrix by a constant value. Since directly computing the necessary offset to make the matrix positive definite is not feasible for such large matrices, the proposed solution consists in iteratively increasing the diagonal offset until the matrix tests positive. It appears clearly in Equation 21 that this only adds a constant factor to the minimized objective function, and thus doesn't affect the optimal solution; however, as acknowledged by (Dhillon-Kulis), by affecting the spread of points and the weight distribution in the matrix, it does affect the operation of algorithms looking for that optimum in unforeseeable ways. **ML : too strong, for a well known problem, this impact can be studied in terms of convergence and quality of the solution and may be positive, so I guess that this unpredictability is only for new type of datasets. MR : toy example : point d'exclamation avec origin au bout du trait**

	min	average $\pm$ variance	max
number of classes	2	8 $\pm$ 9	50
number of time series	56	1626 $\pm$ 2023	9236
time series length	24	372 $\pm$ 400	1882

Table 1: Statistics of the datasets. The length of the times series is expressed in samples.

Such is the price to pay to benefit from the solid geometric foundations and convergence guarantee of the k-means algorithm: a dissociation from immediate semantics, possibly made worse by the necessity of a clunky matrix conditioning procedure.

Our purpose with the work presented in this paper is to propose an alternative algorithm that operates in a way very similar to kernel k-means, but is explicitly designed to process similarities, remains as close as possible to their semantics, and directly attempts to maximize a meaningful quantity: the average intra-class similarity.

## 7 Validation

The claimed features of the k-Average algorithm is now validated on controlled synthetic data.

## 8 Experiments

In order to demonstrate the usefulness of the k-Averages algorithm for realistic data, the clustering of times series is selected as the evaluation task.

Time series, even though represented as vectors and therefore suitable for any kinds of norm-based clustering, is better compared with elastic measures [2, ?]. The Dynamic Time Warping (DTW) measure is an elastic measure widely used in many areas since its introduction for spoken word detection [5] and has been hardly challenged for time series mining [1, 4].

Effective clustering of time series using the DTW measure require similarity based algorithms such as the k-Average algorithm. With some care, kernel based algorithm can also be considered provided that the resulting similarity matrix is converted into a kernel, *i.e.* the matrix is forced to be semi definite positive to be a Gram matrix [3].

### 8.1 Datasets

To compare the proposed algorithm to the kernel k-Means one, a large collection of 43 time series datasets created by many laboratories all over the world and compiled by Prof. Keogh ([www.cs.ucr.edu/~eamonn/time\\_series\\_data](http://www.cs.ucr.edu/~eamonn/time_series_data)) is considered. Statistics about the morphology of those datasets are summarized by Table 1.

In order to ease comparison, the properties and the performance of clustering algorithms under evaluation are broken into 3 tables. Table ?? lists the bi class datasets, *i.e.* the datasets annotated in terms of presence or absence of a given

property. Table ?? lists the datasets with a small number of classes (from 2 to 7) and Table ?? lists the datasets with a larger number of classes (from 8 to 50).

## 8.2 Evaluation Protocol

For every datasets, the training and testing datasets are joined. If relevant, the DTW is computed using the implementation provided by Prof. Ellis (<http://www.ee.columbia.edu/~dpwe/resources/matlab/dtw>) with default parameters.

Clustering is done by requesting a number of cluster equal to the actual number of classes in the dataset. Clustering is repeated 20 times with varying initial conditions, *i.e.* the initial assignment of points to clusters is randomly determined. For fairness of comparison, those assignments are the same for each algorithm.

Several metrics are available to evaluate the performance of a clustering algorithm. The one closest to the actual target application is the raw accuracy, that is the average number of items labelled correctly after an alignment phase of the estimated labelling of the reference one [?].

Another metric of choice, is the Normalized Mutual Information (NMI) criterion. Based on information theoretic principles, it measures the amount of statistical information shared by the random variables representing the predicted cluster distribution and the reference class distribution of the data points. If  $P$  is the random variable denoting the cluster assignments of the points, and  $C$  is the random variable denoting the underlying class labels on the points then the NMI measure is defined as:

$$NMI = \frac{2I(C; K)}{H(C) + H(K)} \quad (24)$$

where  $I(X; Y) = H(X)H(X|Y)$  is the mutual information between the random variables  $X$  and  $Y$ ,  $H(X)$  is the Shannon entropy of  $X$ , and  $H(X|Y)$  is the conditional entropy of  $X$  given  $Y$ . Thanks to the normalization, the metric stays between 0 and 1, 1 indicating perfect matching and can be used to compare clustering with different number of clusters. Random prediction gives an NMI close to 0, whereas the accuracy of a random prediction on a balanced bi class problem is 50%.

For those reasons and ease of reading, only the NMI is considered. Though, it shall be noted that the most of statements that will be made in the following in terms of ranking of the different algorithms still holds while considering the accuracy as reference, see for full results.

			kMeans	kkMeans	kAverages	kAverages	kAverages	kAverages
					object	raw	object	raw
id	nbClasses	nbSamplesPerClass			p	p	b	b
7	2	28±1	48±12	<b>48±32</b>	<b>54±26</b>	<b>66±32</b>	<b>48±26</b>	<b>51±37</b>
12	2	100±47	12±2	14±3	<b>15±0</b>	10±5	<b>15±0</b>	10±5
13	2	442±0	0±0	2±2	<b>4±6</b>	<b>5±2</b>	<b>9±12</b>	<b>5±2</b>
17	2	100±0	0	0	0	<b>1</b>	0	<b>1</b>
20	2	548±1	<b>0±0</b>	<b>1±0</b>	<b>0±0</b>	<b>1±0</b>	<b>1±5</b>	<b>1±0</b>
21	2	60±18	<b>3±0</b>	1±0	1±0	<b>3±4</b>	1±0	<b>4±4</b>
25	2	636±69	30±0	<b>52±0</b>	41±0	36±1	41±0	36±1
28	2	310±54	55±21	<b>78±0</b>	72±1	21±0	72±1	21±0
29	2	490±161	<b>24</b>	21	22	15	22	15
34	2	581±0	0±0	7±0	7±0	<b>8±0</b>	7±0	<b>8±0</b>
42	2	3582±3988	<b>0</b>	0	0	0	0	0
43	2	1650±170	0±0	<b>0±0</b>	<b>0±0</b>	0±0	<b>0±0</b>	0±0

Table 2: nbIterations: 200, distance: dtw, normalizeData: 1, nbRuns: 20

			kMeans	kkMeans	kAverages	kAverages	kAverages	kAverages
					object	raw	object	raw
id	nbClasses	nbSamplesPerClass			p	p	b	b
3	5	12±0	<b>30±4</b>	29±3	<b>32±2</b>	<b>31±4</b>	29±8	20±4
4	3	310±0	36±1	<b>51±4</b>	<b>51±3</b>	42±12	<b>52±3</b>	44±12
5	3	1436±755	0±0	0±0	0±0	<b>0±0</b>	0±0	<b>1±1</b>
6	4	355±0	23±3	<b>44±8</b>	41±5	<b>46±0</b>	40±7	<b>45±4</b>
11	4	80±31	<b>83±3</b>	<b>81±10</b>	65±10	57±16	21±20	32±27
15	4	28±5	45±4	67±9	<b>72±10</b>	63±3	<b>73±7</b>	65±7
18	5	93±9	9±0	<b>9±1</b>	<b>10±1</b>	8±1	<b>9±2</b>	8±1
19	7	93±22	<b>5±1</b>	<b>5±1</b>	<b>5±0</b>	<b>5±1</b>	<b>5±0</b>	5±1
22	7	20±8	44±2	51±4	50±2	<b>54±1</b>	44±15	37±15
26	6	74±20	22±3	23±3	<b>25±2</b>	21±1	<b>24±2</b>	21±2
27	4	15±8	66±4	59±9	61±7	<b>72±3</b>	53±19	<b>70±11</b>
30	3	3079±2045	<b>60±0</b>	<b>61±4</b>	<b>60±0</b>	<b>61±0</b>	<b>60±0</b>	<b>61±0</b>
32	6	170±9	76±6	<b>79±4</b>	<b>80±1</b>	77±5	<b>77±18</b>	68±17
33	4	50±0	53±2	<b>58±7</b>	53±3	52±2	<b>56±6</b>	43±15
35	4	1250±43	2±0	<b>15±13</b>	11±8	<b>12±13</b>	<b>16±7</b>	<b>13±12</b>
37	7	50±0	31±2	<b>42±2</b>	<b>42±2</b>	40±2	<b>38±10</b>	17±8
38	6	100±0	79±3	84±4	<b>87±5</b>	<b>87±1</b>	84±5	65±26

Table 3: nbIterations: 200, distance: dtw, normalizeData: 1, nbRuns: 20

			kMeans	kkMeans	kAverages	kAverages	kAverages	kAverages
id	nbClasses	nbSamplesPerClass			object p	raw p	object b	raw b
1	50	18±21	64±1	70±1	71±1	<b>72±1</b>	26±29	12±0
2	37	21±2	58±1	<b>62±1</b>	60±1	58±1	26±25	10±1
8	12	65±0	26±1	29±2	<b>31±1</b>	27±2	<b>29±7</b>	19±9
9	12	65±0	30±1	<b>35±2</b>	<b>35±1</b>	33±1	<b>35±1</b>	24±9
10	12	65±0	25±1	30±1	<b>31±1</b>	28±2	<b>30±7</b>	20±9
14	14	161±73	37±2	<b>77±3</b>	74±1	67±3	74±2	57±17
16	14	161±73	37±2	<b>77±3</b>	<b>77±2</b>	70±3	<b>72±17</b>	52±21
23	8	300±0	87±5	<b>88±5</b>	87±3	<b>90±4</b>	75±33	45±26
24	10	114±171	25±1	<b>32±1</b>	30±2	31±2	28±7	25±8
31	15	75±0	54±2	<b>66±3</b>	<b>67±2</b>	56±2	40±33	10±12
36	25	36±41	42±1	<b>51±1</b>	<b>51±1</b>	51±1	30±25	16±16
39	8	560±0	44±1	<b>46±1</b>	<b>46±0</b>	<b>46±1</b>	<b>46±1</b>	<b>45±5</b>
40	8	560±0	44±0	<b>45±0</b>	44±0	<b>45±1</b>	44±0	<b>44±5</b>
41	8	560±0	0±0	43±1	<b>44±1</b>	42±1	<b>44±0</b>	42±0

Table 4: nbIterations: 200, distance: dtw, normalizeData: 1, nbRuns: 20

### 8.3 Results

### 8.4 From Progressive to Batch Updates

### 8.5 From Raw to Object Normalized Criterion

## 9 Discussion

## 10 Conclusion

## References

- [1] Donald J. Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In Usama M. Fayyad and Ramasamy Uthrusamy, editors, *KDD Workshop*, pages 359–370. AAAI Press, 1994.
- [2] Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. Querying and mining of time series data: Experimental comparison of representations and distance measures. *Proc. VLDB Endow.*, 1(2):1542–1552, August 2008.
- [3] Gert R. G. Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I. Jordan. Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.*, 5:27–72, December 2004.
- [4] Thanawin Rakthanmanon, Bilson Campana, Abdullah Mueen, Gustavo Batista, Brandon Westover, Qiang Zhu, Jesin Zakaria, and Eamonn Keogh. Addressing big data time series: Mining trillions of time series subsequences under dynamic time warping. *ACM Trans. Knowl. Discov. Data*, 7(3):10:1–10:31, September 2013.

- [5] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 26(1):43–49, Feb 1978.

dataSet	id
50words	1
Adiac	2
Beef	3
CBF	4
ChlorineConcentration	5
CinC_ECG_torso	6
Coffee	7
Cricket_X	8
Cricket_Y	9
Cricket_Z	10
DiatomSizeReduction	11
ECG200	12
ECGFiveDays	13
FaceAll	14
FaceFour	15
FacesUCR	16
Gun_Point	17
Haptics	18
InlineSkate	19
ItalyPowerDemand	20
Lighting2	21
Lighting7	22
MALLAT	23
MedicalImages	24
MoteStrain	25
OSULeaf	26
OliveOil	27
SonyAIBORobotSurface	28
SonyAIBORobotSurfaceII	29
StarLightCurves	30
SwedishLeaf	31
Symbols	32
Trace	33
TwoLeadECG	34
Two_Patterns	35
WordsSynonyms	36
fish	37
synthetic_control	38
uWaveGestureLibrary_X	39
uWaveGestureLibrary_Y	40
uWaveGestureLibrary_Z	41
wafer	42
yoga	43

Table 5: normalizeData: 1