**1. Algorithm Name**

[SubspaceKMeans](https://github.com/tetutaro/subspacekmeans)

**2. Reference**

Dominik M, Wei Y, Claudia P, Christian B. (2017). "Towards an Optimal Subspace for K-Means ". [KDD '17](http://www.kdd.org/kdd2017/), Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: Pages 365-373, Ludwig-Maximilians-University Munchen, Munich, Germany

**3. Motivation for the algorithm (or which problems it tries to solve?)**

The interpretation of what K-means algorithm finds becomes increasingly difficult with growing number of dimensions. Even in lower dimensional spaces it is sometimes hard to tell what structure the algorithm has found.

SubKmeans, is a technique, which extends k-means. It finds a clustering partition and simultaneously a transformation which highlights the structure found in the dataset.

**4. Short Description:**

With in each iteration step sub-kmeans rotates the data based on the current cluster partition. Than, it splits the new feature space in to two orthogonal subspaces first subspace represents the clusters and we call it the cluster subspace. The second space is the complementing space and it is assumed to be unimodal of no cluster structure containing only noisy features. The dimensionality of these two subspaces is optimized in the same time. So sub-kmeans rotates and partitions the data such that the clustered space is a permanent cluster structure and contains only relevant features.

**5. Pseudo-Code**

Remarks:

1. **Don’t copy the code/equations as figures – write your own equations using MS Equation Editor**
2. **Provide two pseudo-code: one for build and one for classify**
3. **Use proper indentation + lines number**
4. **Use the following notations:**
5.  - represent the training set which contains *m* instances
6. - represents the set of classes.
7. *I* – represent an inducer (learning algorithm/base learner)
8. *Mt* – a classifier (classification model) that was trained.

**MultiBoost- Building the ensemble**

Input: *S* – a labeled training set

*I* – a based indcuer

*T* – number of iterations

*J* – vector of integers specifying the iteration at which each subcommittee

should terminate.

1. S’ 🡨S with instance weights assigned to be 1.

2. k🡨1

3. FOR t=1 to T

4. IF *Jk*=t THEN

5. reset S’ to random weights drawn from continuous Poisson distribution.

6. standardize S’ to sum to m.

7. k++

8. END IF

9. *Mt=I(S’)*

10. 

11. IF  THEN

12. reset S’ to random weights drawn from continuous Poisson distribution.

13. standardize S’ to sum to m.

14. k++

15. GOTO 9

16. ELSE IF  THEN

17. 

18. reset S’ to random weights drawn from continuous Poisson distribution.

19. standardize S’ to sum to m.

20. k++

21. ELSE

22. 

23. FOR each 

24. IF THEN

25. 

26. ELSE

27. 

28. END IF

29. IF THEN

30. 

31. END IF

32. END FOR

33. END IF

33. END FOR

**MultiBoost- Classify an instance**

Input: *x* – an instance needed to be labeled

1. Return 

Figure 1: The MultiBoost Pseudo Code

**5+6. Pseudo-Code and Algorithm Explanation:**

|  |  |
| --- | --- |
| **Symbol** | **Interpretation** |
|  | Dimensionality of original space |
|  | Dimensionality of the clustered space |
|  | Number of Clusters |
|  | Set of all objects |
|  | Set of objects assigned to cluster i |
|  | A data point or object |
|  | Dataset mean in the original space |
|  | Mean of cluster i in the original space |
|  | Scatter matrix of the dataset in the original space |
|  | Scatter matrix of cluster i in the original space |
|  | . We perform an eigenvalue decomposition of Σ and use the eigenvectors—sorted ascending to their corresponding eigenvalue—as columns in V . |
|  | Projection onto the first m attributes |
|  | Projection onto the last d −m attributes |
|  | (orthonormal) matrix of a rigid transformation |
|  | l × l identity matrix |
|  | l × r zero matrix |

|  |  |  |
| --- | --- | --- |
| 1 | |  |
| Explanation | The input of the subkmeans algorithm is: 1) the dataset D  2) the number of clusters K | |
| 2 | |  |
| Explanation | | The output of subkmeans is:  1) k clusters  2) the rotation matrix (which is the eigenvector matrix)  3) the dimensionality of the clustered space |
| 3 | |  |
| Explanation | | Initializing a random orthonormal matrix. That means: . |
| 4 | |  |
| Explanation | | Initializing the number of dimensions to be used in a cluster after dimension reduction process. This value will be changed and converged. |
| 5 | |  |
| Explanation | | Calculating the mean vector of all data set D. |
| 6 | |  |
| Explanation | | Calculating the Scatter matrix which is used to make [estimates](https://en.wikipedia.org/wiki/Estimation_of_covariance_matrices) of the [covariance matrix](https://en.wikipedia.org/wiki/Covariance_matrix). The covariance matrix is used in the eigendecomposition process. |
| 7 | |  |
| Explanation | | For each K, which will become a cluster we choose a random point that based upon her the cluster will start building itself. |
| 8 | | **While (**number of dimensions decreases): |
| Explanation | | Start while loop that stops when the number of dimensions is stops decreasing after dimension reduction of the clusters. |
|  | | //Assignment step |
| 8 | |  |
| Explanation | | Assign to every cluster an empty group. |
| 9 | |  |
| 10 | |  |
| 11 | |  |
| Explanation | | Assign every instance to a cluster. Do it by finding the cluster that is the closest to the instance based on the transformed dimensions. Using calculations of eigendecomposition projection differences of X and the cluster mean point using the first m eigenvalues. |
|  | | //Update step |
| 12 | |  |
| Explanation | | **For** I in 1:k (for every cluster) |
| 13 | |  |
| Explanation | | Evaluate the mean of the I’th cluster |
| 14 | |  |
| Explanation | | Evaluate the scatter matrix/covariance matrix of the I’th cluster |
|  | | //Eigendecomposition: |
|  | | // Ɛ: list if eigenvalues in ascending order |
|  | | // V: corresponding eigenvector |
| 15 | |  |
| Explanation | | Eigen decomposition – calculate eigenvectors and eigenvalues |
| 16 | |  |
| Explanation | | Setting m to the number of negative eigenvalues of Σ |
|  | | **End for** |
|  | | **End while** |
|  | |  |

**7. Illustration**

In this section we illustrate the first two iterations of our sub-kmeans algorithm. The only parameter to be tuned is k. We will be using k = 3.

Provide a figure that explains simply what the algorithm does (can do it using the paper or youtube video).

To understand the next example of two sub-kmeans iterations, we will follow the next parameters in each iteration to understand the process:

|  |  |  |  |
| --- | --- | --- | --- |
| Iteration K |  |  |  |
| 0 |  |  |  |
| 1 |  |  |  |
| 2 |  |  |  |

Explanation of the why values change – it is an explanation of k-means

Here is the important part, we will show the Ɛ list and m value in each iteration:

Iteration 0:

m=

Ɛ:

Explanation:

Figure of the cluster:

Iteration 1:

m=

Ɛ:

Explanation:

Figure of the cluster:

Iteration 2:

m=

Ɛ:

Explanation:

Figure of the cluster:

Add a snapshot of the data set we will be using with a simple description

**8. Strengths**

1. *“The non-optimized version is easy to implement and only uses standard features provided by all data computing frameworks.” – It is very simple to implement since it uses no additional parameters in order to receive the non-optimized clustering result.*
2. *“it is fast, though this could most likely further increased – The “time consumer” is the calculation of the covariance matrix, which is an o(n^2 ). Means, even when a very large dataset, the time of processing is quite fast.*

**9. Drawbacks**

1. *“It should be noted that this value may not be chosen too big or too small, as it balances the probability that the initial clustered space contains parts of the cluster structure we want to find and the expressiveness of the similarity measure. For simplicity, we use in our implementation d/2 as it depends on d. Yet, it is also possible to let the user choose an initial value. The optimal value form is subsequently found during the optimization.” –* Since they choose m without reference to the eigenvalues before starting the iterations, the convergence process might be very long. Instead, they should have chosen the m value using the eigenvalues.
2. *“Last but not least, the initial cluster centers could for example be picked randomly from the dataset or set by k-means++”. –* Again, it is inefficient to choose center points randomly. There are better ways to choose the initial centroids such in k-means++.
3. *“Future efforts may be directed towards the approximation of the transformation matrix V. We are very confident that this can be achieved quite easily by using a randomized singular value decomposition, resulting in further performance improvements for very high-dimensional datasets.” – The convergence process can get a lot faster when the transformation matrix V is “decomposed”, especially when we have high dimensional dataset.*

**10. Experimental Results**

Same K

Measures:

* need to choose ones, since K does not change, I believe simple accuracy will do
* Execution time for each iteration
* Number of iterations till convergence

Baseline classifiers: K-means: Will be compared using the same measure of accuracy

10 data sets – say a few words about them

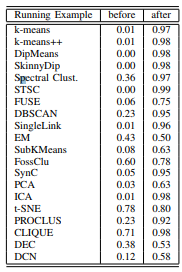
Lots of figures

**11. Conclusions**

What are the conclusions from the results?

**12. Citations**

The paper was cited in 4 other papers:

1. Jang, H. J., Kim, B., Kim, J., & Jung, S. Y. (2018). An Efficient Grid-Based K Prototypes Algorithm for Sustainable Decision-Making on Spatial Objects. Sustainability, 10(8), 2614.‏ - They used sub k-means to reduce complexity of the basic k-means algorithm
2. Schelling, B., & Plant, C. (2018, November). DipTransformation: Enhancing the Structure of a Dataset and thereby improving Clustering. In 2018 IEEE International Conference on Data Mining (ICDM) (pp. 407-416). IEEE.‏ - They compared different techniques that intend reducing dimensionality with the goal of finding a subspace compatible with k-means. The results showed Sub-k-means is not the optimal algorithm:   
   
3. Wang, J., Hsieh, C. J., & Shi, D. (2018, May). NLRR++: Scalable Subspace Clustering via Non-Convex Block Coordinate Descent. In Proceedings of the 2018 SIAM International Conference on Data Mining (pp. 28-36). Society for Industrial and Applied Mathematics.‏ - They mentioned the sub-kmeans clustering method as a past decade method with some other subspace clustering methods.
4. Mautz, D., Ye, W., Plant, C., & Böhm, C. (2018, July). Discovering Non-Redundant K-means Clusterings in Optimal Subspaces. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 1973-1982). ACM.‏ - “Slightly different to the above-described methods is the recently proposed algorithm SubKmeans. It aims to combine the kmeans algorithm with a simultaneous dimensionality reduction step to find one arbitrarily oriented subspace. It is a special instance of Nr-Kmeans, in which we only assume a single clustered space with a single clustering and a complementing noise space.”