Clustering High-dimensional data with Size Constraints Using Heuristic Methods

Research Proposal

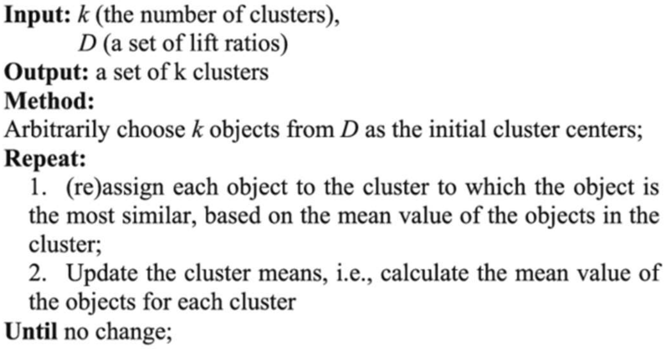
Z. Xu, Postgraduate Student, School of Computer Science, University of Birmingham

I. INTRODUCTION

With the advances in research and applications of machine learning and intelligent data analysis, the amount of data required in these events have soared dramatically in recent years. Machine learning generates automatic approaches that improve decision making while learning from datasets. One class of the machine learning approaches with more significance in applications is supervised learning, which requires class labels of objects as prior identifiers. (Jain, 2010) Supervised learning introduces more control on the dataset and yields good performance even against a big dataset. Nevertheless, the extra information needed to classify the training objects, which is not intrinsic to the dataset, may cost large amount of human efforts to provide a better result of the experiment. In fact, a large percent of the raw data acquired are unlabelled, which makes it inapplicable to supervised learning without labelling.

In real world applications, cluster analysis is applied to process data without external label information. Cluster analysis is defined as the study of approaches to group or cluster objects with respect to their intrinsic similarity or characteristics. (Šárka BrodinováEmail et al., 2019)The goal of data clustering is to find the natural groupings of objects or patterns by partitioning data into a certain number of categories so that objects in the same category are like each other and differ from those in other groups. (Xu, R., & Wunsch, 2005) (ELKI, 2014) Clustering has been considered as the most popular unsupervised learning approach since k-means was introduced in 60s.

Despite the recent development in clustering methods, k-means remains one of the simplest and most famous clustering algorithms. (Jain, 2010) The task is to find clusters of data given by an input of cluster number K. After initiation, it first assigns the data points to the nearest clusters. For each cluster, the centroid is then calculated and updated by the mean of objects. Iteratively, clusters are updated with refinement of the location of each centroid until convergence is met. (ELKI, 2014) Other clustering algorithms also include hierarchical clustering, mixture densities, fuzzy clustering, kernel methods for clustering, neural network-based clustering, etc. In some specific cases, semi-supervised learning is applied to utilise both labelled and unlabelled data for better performance.

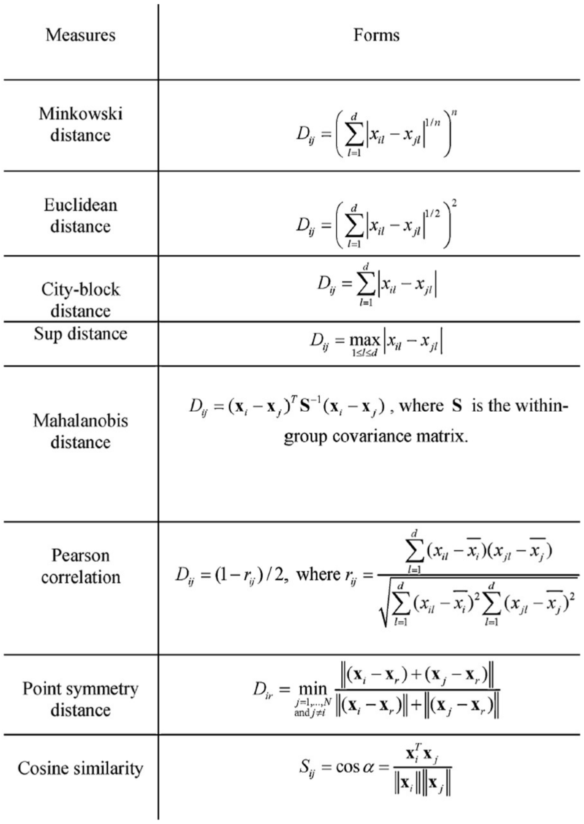


```
Input: k (the number of clusters),
       D (a set of lift ratios)
Output: a set of k clusters
Method:
Arbitrarily choose k objects from D as the initial cluster centers;
Repeat:
  1. (re)assign each object to the cluster to which the object is
     the most similar, based on the mean value of the objects in the
     cluster;
  2. Update the cluster means, i.e., calculate the mean value of
     the objects for each cluster
Until no change;
```

Figure 1- K-Means Algorithm

By applying clustering algorithms, some assumptions are made to understand the underlying structure of data space. (MIT, 2006) It is assumed that points in the neighbourhood are more likely to have the same label, whereas points in the same category are more likely to share the cluster with higher density in the data space. (Peikari, 2018) The assumptions suggest that the similarity among points can be measured by the distance between a pair of objects or a cluster and an object. To evaluate the results of clustering algorithms, distance measures are pivotal in showing either how similar or dissimilar a pair of points are in a quantitative way. As shown in table 1, the most popular metric in clustering is Euclidean distance, which is derived from Minkowski metric with a special case at n = 2. (Xu, R., & Wunsch, 2005) Another commonly used metric is called city-block distance (or Manhattan distance) with linear output. It only counts the sum of distance on block edges. It can be illustrated that higher polynomial metrics are not very interpretable for similarity because variances in these results are dominant, except for sup distance, which takes the maximum of distance in all direction. Pearson correlation is also used to explore the relationship between objects. Clustering analysis relies on adequate amount of data and the suitable metric for the dataset.

Despite the use of analysing similarity, clustering algorithms are less competitive particularly in real-world high-dimensional datasets. The common problem for analysis in high dimension though is caused by “curse of dimensionality”, which drastically increases the computational time and memory usage of the algorithm. As shown in figure 2, the performance of distance metrics between pairs of objects in the data space are adversely disturbed by the high dimensionality because it shows little difference between close and distant pairs in terms of distance, which weakens our assumption to evaluate their similarity. (Elhamifar, E. et al., 2013) The challenges for identifying the groups among high-dimensional data divulge in other aspects as well. The presence of noisy data and outliers make clustering harder to confine in a data space with reasonable size. (Šárka BrodinováEmail et al., 2019) Hence, a more robust clustering method to locate the cluster centroid is required to mitigate the interference from small number of variances or outliers.



Measures	Forms
Minkowski distance	$D_{ij} = \left(\sum_{l=1}^{d} \lvert x_{il} - x_{jl} \rvert^{1/n} \right)^{n}$
Euclidean distance	$D_{ij} = \left(\sum_{l=1}^{d} \lvert x_{il} - x_{jl} \rvert^{1/2} \right)^{2}$
City-block distance	$D_{ij} = \sum_{l=1}^{d} \lvert x_{il} - x_{jl} \rvert$
Sup distance	$D_{ij} = \max_{1 \le l \le d} \lvert x_{il} - x_{jl} \rvert$
Mahalanobis distance	$D_{ij} = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j)$, where $\mathbf{S}$ is the within-group covariance matrix.
Pearson correlation	$D_{ij} = (1 - r_{ij})/2$, where $r_{ij} = \frac{\sum_{l=1}^{d} (x_{il} - \overline{x}_i)(x_{jl} - \overline{x}_j)}{\sqrt{\sum_{l=1}^{d} (x_{il} - \overline{x}_i)^2 \sum_{l=1}^{d} (x_{jl} - \overline{x}_j)^2}}$
Point symmetry distance	$D_{ir} = \min_{j=1,\ldots,N \text{ and } j \ne i} \frac{\lVert (\mathbf{x}_i - \mathbf{x}_r) + (\mathbf{x}_j - \mathbf{x}_r) \rVert}{\lVert (\mathbf{x}_i - \mathbf{x}_r) \rVert + \lVert (\mathbf{x}_j - \mathbf{x}_r) \rVert}$
Cosine similarity	$S_{ij} = \cos\alpha = \frac{\mathbf{x}_i^T \mathbf{x}_j}{\lVert \mathbf{x}_i \rVert \lVert \mathbf{x}_j \rVert}$

Figure 2- Distance Measures

In specific applications of clustering algorithms, people can collect auxiliary information of the size distribution of clusters and pairwise relationships within the dataset. An additional requirement on the clustering results can be described to find the distribution of size of each cluster that fits the prior best. In such cases, the most important goal is to recover underlying low-dimensional subspaces that the result retains some local structures from the hyperspace. (Elhamifar, E. et al., 2013) As a result, the original clustering models are not possible to iteratively process the data and group objects in arbitrary clusters with no regard to their size. It is perceived that a heuristic method is needed to solve the problem (NP hard) from an optimisation perspective. This research is intended to evaluate performance of different approaches to the high-dimensional clustering problem and find solutions that yield best efficiency and effectiveness to fit the size constraints of clusters. In addition, we want to discuss the necessity and effectiveness of dimensionality reduction methods in the scenario of clustering high-dimensional data with size constraints.

$$\lim_{n\to\infty} \mathbb{E}\left(\frac{d_{\max} - d_{\min}}{d_{\min}}\right) \to 0$$

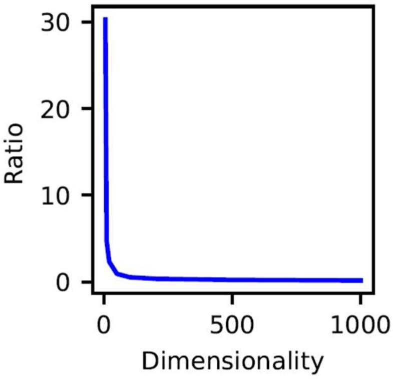


Figure 3 - Curse of Dimensionality (Styles, 2019)

II. LITERATURE REVIEW

In the first phase of project, we propose to study the properties of high-dimensional data and evaluate the performance of different methods to cluster high-dimensional data with constraints. With reference to past works on this subject, there are three main classes of methods to be evaluated. 1) To use iterative methods with heuristic inference of the optimisation. An improved version of k-means is made available to cluster the objects into same-size clusters. (ELKI, 2014)2) Use linear algebra to factorise the problem and solve the optimisation separately; 3) Apply probabilistic methods so that a multivariate distribution of the classes can fit the optimal result of prior constraints. (al., 2004)

Since finding the solution to the proposed problem, clustering with size constraints, is in general NP-hard, it is perceived that the problem can be transformed to an optimisation problem over minimisation by relaxing the convex. (Elhamifar, E. et al., 2013) Elhamifar et al. (Elhamifar, E. et al., 2013) incorporated a solution to sparse subspace clustering problem by a Sparse Subspace Clustering algorithm, which relaxes on convex and optimises over the program. Vali (Vali, 2013) proposed a new optimisation approach called Clustering-Based Parallel Genetic Algorithm (CBPGA) to solve global optimisation problems. Another idea is to implement iterative methods whereas ensure that the constraints are satisfied in each iteration. (Wagstaff, et al., 2001)

In addition, to study the effect of dimensionality reduction methods on the experiment, a collection of different approaches is included to validate the robustness of proposed algorithms. (Scikit-Learn, 2018)

III. Research Questions

As discussed in the sections above, this research project is designed to discuss the following questions:

1. What additional metrics of clustering high-dimensional data are required? Why?
2. By comparing the results, which existing method is the best model in grouping high-dimensional data into clusters of same size?
3. To what extent, a heuristic method can improve the clustering over size constraints?
4. How to apply relaxation to the size constraints?
5. What are the effects of dimensionality reduction on optimisation in the experiment?
6. Can we generalise the algorithm to similar problems? How?
7. Is the project/ experiment successful? Why?

IV. Project Plan

Due to the time limit of the project, to reach the goals proposed as planned, it is better partitioned into several phases.

Week	Goals
Wk. 1 – 2	Scope Research
Wk. 2	Proposal Redraft
Wk. 3	Data Collection & Pre-processing
Wk. 3 – 5	Methods and Algorithms Selection
Wk. 6	Experiment Design
Wk. 6 -7	Experiment Data Analysis
Wk. 8	Inference and Evaluation
Wk. 9 – 10	First Draft Report
Wk. 10 - 11	Iteration and Validation
Wk. 11	Final Draft Report
Wk. 11	Presentation

V. References

1. Elhamifar, E., & Vidal, R. (2013). Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11), 2765-2781.

2. ELKI. (2014). *Same-size k-Means Variation*. Retrieved from Same-size k-Means: https://elki-project.github.io/tutorial/same-size_k_means

3. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, 31(8), 651-666.

4. MIT. (2006). *semi-supervised learning.*

5. Peikari, M. (2018). *Scientific reports*, 8(1), 7193.

6. Šárka BrodinováEmail et al. (2019). *Robust and sparse k-means clustering for high-dimensional data.*

7. Scikit-Learn. (2018). *Dimensionality reduction*. Retrieved from Dimensionality reduction: https://scikit-learn.org/stable/modules/decomposition.html#decompositions

8. Vali, M. (2013). New Optimization Approach Using Clustering-Based Parallel Genetic Algorithm. *arXiv preprint*, 1307.5667.

9. Wagstaff, et al. (2001). Constrained k-means clustering with background knowledge. In Icml (Vol. 1, pp. 577-584).

10. Xu, R., & Wunsch. (2005). *Survey of clustering algorithms.* D.C.

11. Y. Sugaya and K. Kanatani, “Geometric structure of degeneracy for multibody motion segmentation,” in Workshop on Statistical Methods in Video Processing, 2004.