International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No. (02), January 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

Improved Online k-Means Algorithm

Divya Dadhich[1], Dr. Amit Sharma[2]

[1] M.Tech Scholar, [2]Professor, Department of Computer Science & Engineering,
Vedant College of Engineering & Technology, Bundi, Rajasthan (India)

ABSTRACT

A restricted computation model of Online clustering in which data points arrive one by one and clustering decisions can neither be postponed or reconsidered is a non-trivial problem to solve. Associating it with objective function of the popular k-means algorithm gives an insight into behavior of clustering method and its application to the online clustering. This paper improved upon an online k-means algorithm such that the output cluster pattern can be evaluated according to value of objective function.

Keywords: Clustering, K-means, Offline Data, Online Data, Streaming Data

I. INTRODUCTION

Clustering is a well-studied problem domain of machine learning with its vast applications in data analyses, image processing, pattern recognition, the clustering techniques are equally sought after by business analysts, scientists and information engineers. Clustering problem can have broadly three types of settings: standard offline setting, streaming model and online clustering. In the standard offline model [1, 2,3] entire data is known a priori to clustering, hence the clustering decisions can be taken very precisely according to well-formulated objective functions. Streaming model [4,5,6] allows a single pass through the data with limited memory. The clustering decision are output finally when the stream is over. Online clustering [7,8,9] has a model in which the clustering decisions are made as and when data points arrive, which arrive one by one and arbitrarily. This restricted model of online clustering is now most sought after by researchers since it suits many ad-hoc data analysis needs of modern applications.

As the technology develops, much amount of data is produced and injected as a stream for online processing make it impossible for the conventional clustering method to be useful. Such applications require fast yet effective methods for grouping data. Achieving sufficient clustering quality within stimulated time is the major requirement.

The major approach towards online clustering is to extend the existing offline algorithms for the online problem. The k-centres algorithm proposed by Charikar et al [7] uses an incremental approach. Online version of Expectation-Minimization is attempted by [8]. Formally provable results of k-means as an online clustering method are proposed in work by Choromanska and Montoloni [9]. Liberty et al[10] also use k-means as basic clustering algorithm but their method is very different from the actual method of k-means. The concept of facility location problem [11] is used to formulate the online clustering problem.

This paper proposes an improvement of the work by Liberty et al. The online clustering algorithm proposed in [10] is based on the classic k-means technique yet very few concepts of the conventional k-means are used. This

paper proposes to use the objective function of k-means clustering problem and adapt it as the cost of opening a new cluster for the incoming point of data stream. Besides modifying the cost function, the proposed algorithm has an added phase called merging phase which is executed once all data has arrived. The proposed process is linear in size of stream and produces a good quality cluster structure.

The paper can be organized as follows. Section II discusses the proposed online clustering algorithm and the development of idea behind the proposal. Section III discusses the experimental results on some synthetically generated and real-life datasets. Section IV compares the proposal with the work of Liberty et al in terms of cost.

II. PROPOSED ONLINE CLUSTERING ALGORITHM

2.1 Liberty et al's contribution

Liberty, Sriharsha and Sviridenko [10] proposed the online k-means algorithm. The formal algorithm is listed in Fig. 1.

Algorithm: Online k-means algorithm
Input: V, k
Output: Atleast k+1 cluster formation
Step 1: $C \leftarrow$the first k+1 distinct vectors in V; and n=k+1
Step 2: For each of these yield itself as its cluster
Step 3: $w^* \leftarrow min_{v,v' \in C} \|v - v'\|^2/2$
Step 4: $r \leftarrow 1; q_1 \leftarrow 0, f_1 = {w^*}/{k}$
Step 5: for $v \in$ the remainder of V do
Step 6: $n \leftarrow n + 1$
Step 7: with probability $p = min(D^2(v,C)/f_r, 1)$
Step 8: $C \leftarrow C \cup \{v\}, q_r \leftarrow q_r + 1$
Step 9: if $q_r \geq 3k(1 + log(n))$, then
Step 10: $r \leftarrow r + 1, q_r \leftarrow 0; f_r \leftarrow 2 \cdot f_r - 1$
Step 11: end if
Step 12: yield: $c = \text{argmin}_{c \in C} \|v - c\|^2$
Step 13: End for

Fig. 1 Online k-means algorithm

The online k-means algorithm aims to bifurcate an online arriving stream of data V into relevant clusters. v denotes a data point in the entire dataset V whose value is not known since it is an online stream. The algorithm takes as input a parameter k that denotes the minimum number of clusters that will be formed as output of the algorithm. Hence, atleast k+1 clusters are formed. The initial points n arriving from the online stream of data, equal to k+1, are assigned as the initial k+1 cluster centers in Step 1. These k+ 1 points constitute the cluster center set C. The algorithm is not an iterative algorithm, rather conducts in r phases. For r^{th} phase, f_r denotes the facility opening cost, or precisely, the cost of opening a new cluster and q_r denotes the number of clusters in r^{th} phase. The distance between each of the points in C is calculated and the square of the minimum distance divided by 2k is the initial facility opening cost as explained through steps 3 and 4 of the algorithm. Since no proper clusters are formed, the value of q_r remains nil. New cluster formation is opened with the probability p

which is the actually the minimum squared distance between the entering point v and the existing cluster centers divided by the cost f_r of that phase. With each successive phase, cost f_r is double of the previous cost so as to ensure that lesser clusters are opened in later phases. The algorithm moves to the next phase when $q_r \geq 3k(1 + log(n))$.

2.2 Development of Idea

The Online k-means algorithm is observed to be lacking in two aspects.

- Though it has been named as k-means, the 'means' are never used. This would clearly affect the overall cluster structure obtained.
- The resultant cluster structure is expected to contain a high noise component. The reason behind this is that some of the initially arriving points may not attract any of the points arriving later in the stream into their clusters. These points can then form very small or singleton clusters.

In the dissertation, we propose the following improvements related to the above discussed limitations.

- The first limitation is dealt with by updating centroids at every phase of the algorithm. Updation is done phase-wise and not at every point because it will add to the time complexity of the algorithm.
- The second limitation is solved by introducing a merging phase in the algorithm after the entire clustering process is over. For this, clusters with very low population are identified and merged with nearest big cluster.

For a subset S_i, if

$$|S_i| < \frac{N}{10}$$

Then, subset S_i of V is a noise cluster and hence has to be merged. N refers to the number of data points in the entire online stream.

2.3 Proposed Improved Online K-means

The discussed improvements can then be incorporated into the online k-means algorithm. The complete description of the proposed Improved Online K-Means algorithm is given in Fig. 2.

International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No. (02), January 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

Algorithm: Improved Online k-means algorithm
Input: Online stream of data V, parameter k

Step 1: Assign number of data points arriving from the online stream $n = k + 1$ as initial cluster centers forming a cluster center set C.
Step 2: For each of these yield itself as its cluster
Step 3: For first phase or $r \leftarrow 1$, calculate facility opening cost as $f_1 = {w^*}/{k}$, $w^* \leftarrow min_{v,v' \in C} \|v - v'\|^2/2$
Step 4: Number of clusters in first phase, $q_1 \leftarrow 0$,
Step 5: for $v \in$ the remainder of online data stream V do
Step 6: $n \leftarrow n + 1$
Step 7: with probability $p = min(D^2(v, C)/f_r, 1)$
Step 8: $C \leftarrow C \cup \{v\}, q_r \leftarrow q_r + 1$
Step 9: if $q_r \geq 3k(1 + log(n))$, then
Step 10: Change the phase as $r \leftarrow r + 1$
Step 11: Set phase parameters $q_r \leftarrow 0; f_r \leftarrow 2 \cdot f_r - 1$
Step 12: Update the centroids C as means of clusters
Step 12: end if
Step 13: yield: $c = \text{argmin}_{c \in C} \|v - c\|^2$
Step 14: End for
Step 15: For all subsets S_i with $|S_i| < \frac{N}{10}$, merge into nearest big subset.

Fig. 2 Online k-means algorithm

III. EXPERIMENTAL RESULTS

3.1 Experimental Setup

The proposed Improved Online K-means algorithm is tested on various synthetically generated and real life datasets using the MATLAB computing platform. The complete description of the synthetic and real-life datasets is provided in Tables 1 and 2.

TABLE 1

DESCRIPTION OF THE SYNTHETICALLY GENERATED DATASET

Dataset	Number of instances	Number of dimensions	Desired Number of classes
A1	3000	2	20
A2	5250	2	35
A3	7500	2	50
S1	5000	2	15
S2	5000	2	15
S3	5000	2	15
S4	5000	2	15
D31	3100	2	31
R15	600	2	15

TABLE 2

DESCRIPTION OF REAL LIFE DATASET

Dataset	Number of instances	Number of dimensions	Desired Number of classes
House(5 bits per color)	34112	3	256
Bridge	4096	16	256
Shuttle	58000	9	7

3.2 Criteria for evaluation

The performance evaluation of proposed work is done taking the following criteria.

1. ***Ratio of output to desired number of clusters:*** The lesser the ratio, more close the output is towards the desired output. This criterion can be considered only for evaluating an algorithm against the datasets for which ground truth clusters are known. It is better to use this ratio instead of accuracy in context of online clustering when the application may always have a user-defined input for desired number of clusters.
2. ***Cost:*** The proposed algorithm updates the position of centroids at every phase. Once all the data has arrived, the centroids are again updated and small clusters are merged into bigger clusters. The effect of these two changes in the Online K-means algorithm by Liberty et al can thus be observed through value of objective function before the merging of clusters and after merging. Cost of liberty et al's work is also considered and is entirely different from the cost before and after merging of the proposed algorithm.

   In general, a high value of objective function indicates poorer quality of clusters. Reason of high value of the objective function is that when the output pattern has large sparse clusters, the cost will be high. If, on the other hand, the output cluster has small dense clusters, the cost will be low.

   When talking about a non-general case, output clusters are very less in number before merging and in online k-means. In the proposed algorithm however, none of the data points have been left unclustered so the noise becomes a part of the existing cluster. But the cluster quality is not poor because the number of output clusters after merging is very close to the number of desirable clusters.

IV. RESULTS ON SYNTHETIC DATASETS

The performance evaluation of the proposed clustering is done with respect to the cost of the algorithm before and after the merging phase and ratio of output to desired number of clusters. Cost is the value of the objective function of SSE of k-means. Experimental results in this section cover the discussed synthetic datasets. The desired number of clusters is varied from 30 to 50 at an interval of 5 for evaluation. Fig. 3 (a)(b)(c) depict results for the ratio of desired to output clusters, cost before merging and cost after merging.

With increasing value of the desired number of clusters, a significant variation in the ratio and cost is observed. The reasons for such variations are large number of cluster formations and correspondingly high rate of merging data points including noise components. However, in each case, with increase of the desired number of clusters, a drop in value of ratio and cost towards a constant value is obtained. This means that the algorithm is returning results as desired consequently. Robustness of the algorithm is thus guaranteed.

International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No. (02), January 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

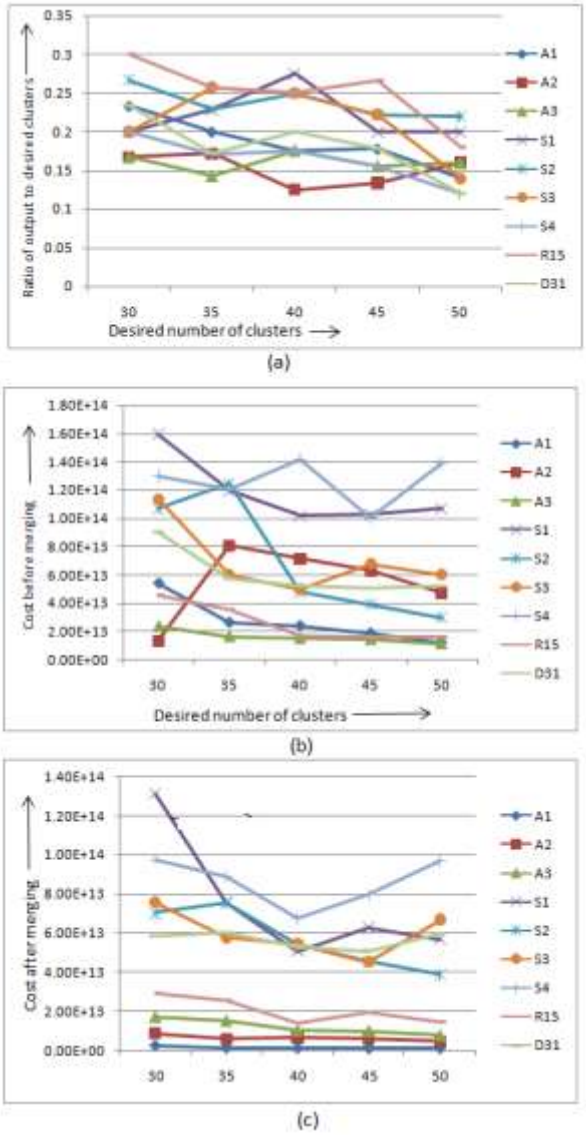


Fig 3 Results on synthetic datasets

V. RESULTS ON REAL-LIFE DATASETS

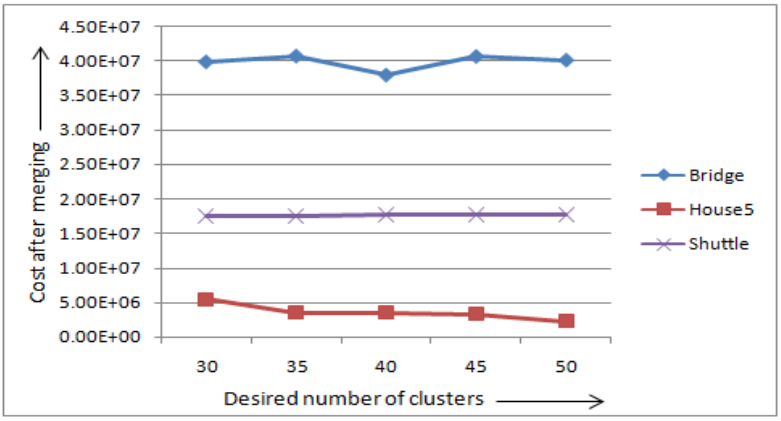
The performance of the proposed algorithm is next evaluated on real-life datasets, House, Bridge and Shuttle. The parameter varied for experiments is the desired number of clusters from 30 to 50. Fig. 2 illustrates the results.

Results portray variations with increasing number of desired clusters for ratio and cost before merging values of the proposed algorithm, reason being the same as discussed before. However, almost constant values are observed for cost after merging indicating results near to optimal by the algorithm. Consistency in results again denotes a robust performance of the proposed algorithm.

International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No. (02), January 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

(a)

(b)



(c)

Fig 4 Results on Real-Life Datasets

VI. COMPARISON RESULTS

The proposed algorithm is compared with Liberty et al's work[10] in terms of the cost of the algorithms. The comparison is done dataset-wise for all the synthetic and real life datasets. Cost 1 in the experiments denotes the cost of Liberty et al's work. Cost 2 and Cost 3 refers to the cost of the proposed algorithm before and after the merging phases respectively. Fig.5 depicts the comparison results for all synthetic datasets. Fig. 6 illustrate results of comparison between the proposed and Online k-means algorithm on real life datasets.

International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No. (02), January 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

Taking the general point of view, a reduced cost indicates better performance. The same is illustrated on all synthetic datasets. The cost after merging is very less compared to the other two costs. No two costs are similar for any dataset. An even better performance is observed with increasing desired number of clusters.

The results on Bridge and House dataset are as desired favoring the proposed algorithm. A much higher cost after merging in case of Shuttle Dataset indicates cost due to merging all data points including noise components. However, a uniform proportion with increasing desired number of clusters shows consistency and robustness of the proposed algorithm. It further means that the quality of the obtained clusters using the proposed algorithm is not poor.

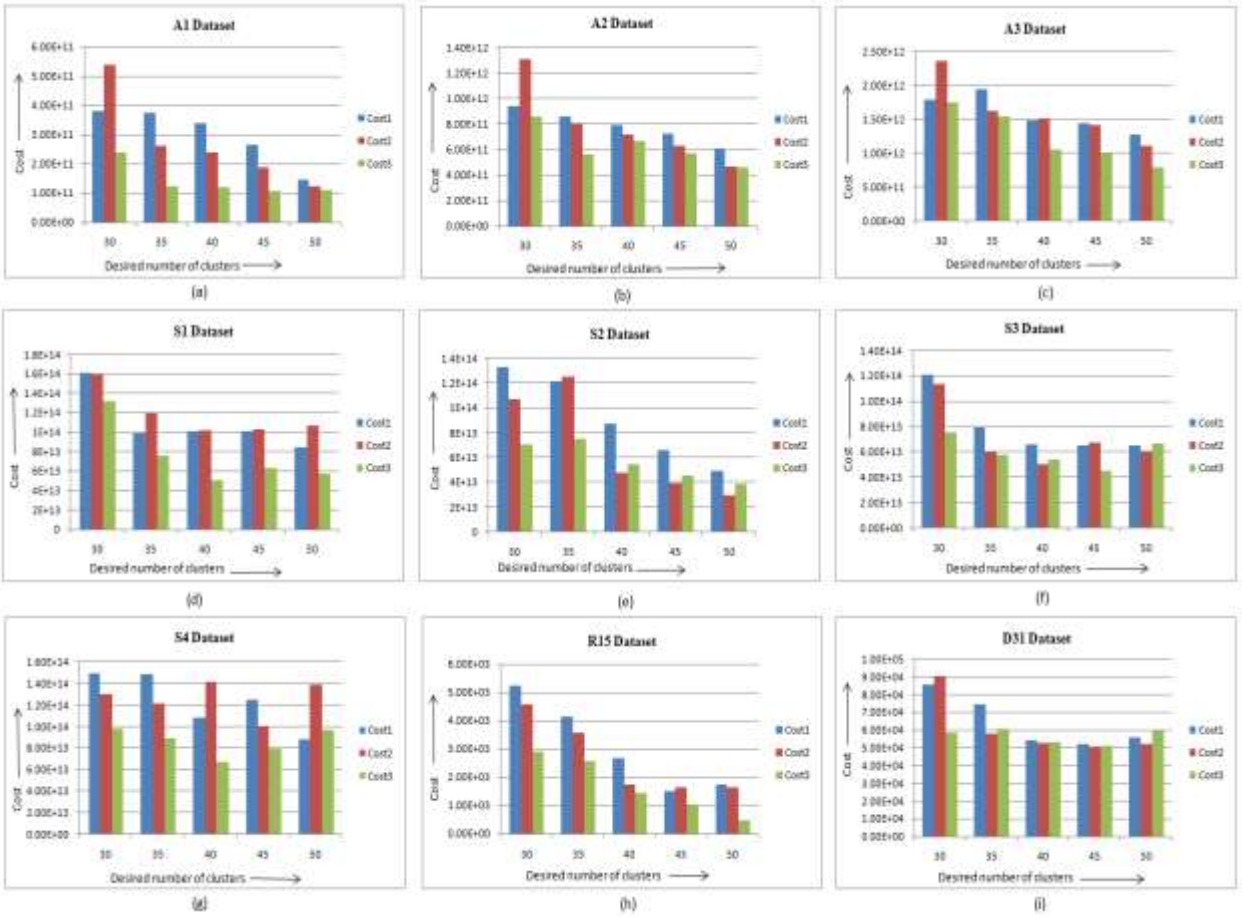


Fig 5. Results on synthetic datasets (a) A1, (b) A2, (c) A3, (d) S1, (e) S2, (f) S3, (g) S4, (h) R15, (c) D31

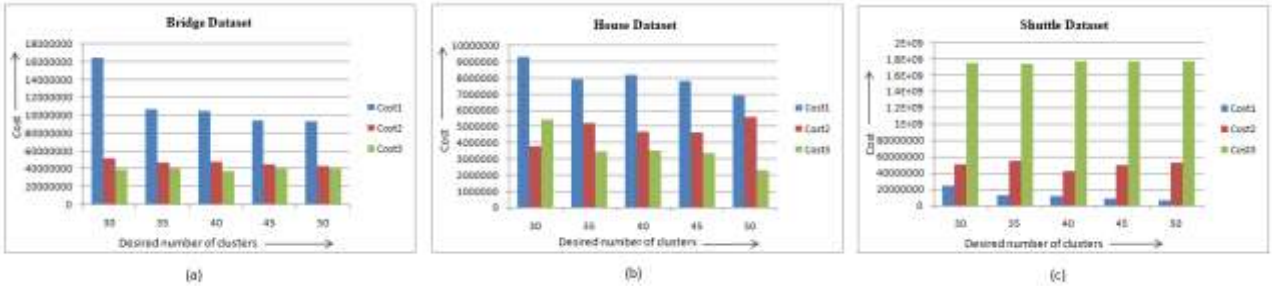


Fig 6 Results on real life datasets (a) Bridge, (b) House (c) Shuttle

VII. CONCLUSION

Online clustering is a non-trivial problem of clustering arbitrary arriving data within a very restricted model. the conventional clustering algorithms that deal with standard offline setting cannot properly deal with online model and simple extensions are not possible. Rather clustering methods for online clustering are to be developed with proper problem formulation and modification to the standard methods.

This paper presents an online clustering method based on k-means method in the sense of its objective function formulation. The decision of putting an arriving data point into an existing cluster or creating a new cluster is done similar to facility location problem. Once the cluster decisions have been made, a merging phase picks very small clusters and merges them into the bigger ones. Thus, cluster structure improves and desired number of clusters can be achieved. It doesn't revise all the clusters, rather very small portion of the cluster output is revised. This is in conformance with the restricted memory model.

The behavior of the proposed algorithm according to changing input parameters is studied thoroughly through experiments on popular synthetic and real life datasets. The proposed method shows its robustness and consistency of output against variation in input parameter, making it behave similar to parameter-free algorithms.

REFERENCES

[1] S. P. Lloyd, "Least squares quantization in pcm", *IEEE Trans. Inf. Theory*, Vol. 28, No. 2, pp. 129–137, 1982.

[2] D. Arthur and S. Vassilvitskii, " k-means++ the advantages of careful seeding". In Nikhil Bansal Kirk Pruhs, and Clifford Stein, editors, *SODA*, pp.1027–1035. SIAM, 2007.

[3] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "A local search approximation al-gorithm for k-means clustering.", *Proceedings of the Symposium on Computational Geometry*, 2002, pp. 10–18.

[4] N. Ailon, R. Jaiswal, and C. Monteleoni. "Streaming k-means approximation", In Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta, editors, *NIPS*, pp. 10–18. Curran Associates, Inc., 2009.

[5] A. Meyerson, M. Shindler, and A. Wong, "Fast and accurate k-means for large datasets", *NIPS*, 2011.

[6] A. Aggarwal, A. Deshpande, and R. Kannan, "Adaptive sampling for k-means clustering", *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, Proceedings of the 12th International Workshop, APPROX 2009, and 13th International Workshop, RANDOM 2009*, 2009, pp.15–28.

[7] M. Charikar, C. Chekuri, T. Feder, and R. Motwani. "Incremental clustering and dynamic information retrieval", *Proceedings of the Twenty-ninth Annual ACM Symposium on Theory of Computing*, STOC '97, 1997, pp. 626–635. ACM.

[8] P. Liang and D. Klein, "Online EM for unsupervised models, *Proceedings of the Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, 2009, pp. 611–619.

[9] A. Choromanska and C. Monteleoni. "Online clustering with experts", *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2012*, 2012, pp. 227–235.

[10] .E. Liberty, R. Sriharsha, and M. Sviridenko, "An Algorithm for Online K-Means Clustering", *Proceedings of the Eighteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, 2016, pp. 81-89.

[11] A. Meyerson, "Online facility location", *FOCS*, pp. 426–431. IEEE Computer Society, 2001.