Achieving Balance in Clusters- A Survey

Divya Saini[1], Manoj Singh[2]

[1] M.Tech Scholar, Gurukul Institute of Technology, Kota, Rajasthan, India
[2] Assistant Professor, Gurukul Institute of Technology, Kota, Rajasthan, India

---***---

Abstract - ***K-means Clustering Algorithm, among the various clustering algorithms proposed till date, has proved its superiority by its simplicity and usability. However, it is prone to a number of limitations, one being lack of balance in clusters. Clusters obtained, if balanced, will result in equally sized clusters thereby distributing the load equally and enhancing the quality of clustering. The existing clustering algorithms aiming at optimizing the traditional k-means are limited to working on the bad initialization problem or the local optimum problem of k-means. This paper discusses some of the relevant research works in the direction of obtaining balance in clusters, its need and approaches.***

Key Words: *K-means, Clustering, Balanced Clustering*

1. INTRODUCTION

Clustering, in data mining, is a process aiming at segmentation of data such that homogeneous data fall in the same group, specifically called a 'cluster' here. Clustering has long been identified and researched upon because of its applicability in various applications like image processing, pattern recognition, data analysis, bioinformatics, machine learning etc. Among the proposed clustering algorithms till date, K-means clustering algorithm [1] has been ranked as one of the top ten clustering algorithms [2]. The reasons for the same lie in its simplicity, easy implementation, flexibility and usability in almost every field. Its popularity and importance can best be felt by the amount of work done on optimizing the algorithm for over 50 years [3]. Optimizations in k-means are done seeing the limitations posed by the algorithm related to bad initialization, convergence to a local optimum value, scalability etc. However, one more issue to deal with is imbalance in formed clusters. In cases of applying k-means to applications not aware of the shapes of clusters, the resultant clustering should aim at identifying high density areas and making the distribution in such a way that each cluster gets equal population or equal density. Instead of grouping, the clustering is an optimization problem in such applications and should be called balanced clustering. Balanced Clustering can be seen in applications of workload balancing, circuit design, image processing etc. This paper provides a brief survey discussing the research headed towards achieving balance in clusters.

2. BALANCING ACCORDING TO NEED

According to Malinen and Franti in [4], there are two following needs/requirements of achieving balance.

- Minimizing Squared Mean Error (SME)
- Balancing cluster sizes

Minimizing SME was the objective of the traditional k-means clustering objective, but it failed to balance cluster sizes. According to the need, the existing balanced clustering algorithms can further be categorized into

- Balance Driven
- Balance Constrained

The two needs specified contradict each other in the traditional k-means algorithm. Clustering, with any of the requirements not met, affects the quality of the clusters formed. In Balance driven algorithms, minimizing SME is the mandatory requirement with balancing cluster sizes as the secondary one. Balance constrained is just the opposite of Balance driven, i.e. balancing sizes of clusters as mandatory. Given below is Table. 1 listing the various Balance driven and Balance constrained clustering algorithms taken from [4].

Balance Constrained	Balance Driven
Balanced k-means[4]	FSCL[7]
Constrained k-means[5]	FSCL-Additive bias[8]
Size constrained [6]	Cluster sampled data[9]
	Ratio cut[10]
	SR-Cut[11]
	Submodular fractional programming[12]

Table -1: List of Balance Constrained and Balance Driven Algorithms [4]

2.1 B*alance Constrained Algorithms*

Balanced k-means [4]- Malinen and Franti contributions to balanced clusters, apart from describing the need and categorization of the existing algorithms, involve changes in the assignment step of the traditional k-means clustering. The k-means clustering consists of teo main steps; the assignment step and the update step, the former conventionally assigning the data points to the selected nearest centroids and the latter calculating the mean of each cluster. The concept of the assignment is replaced by introducing n pre-allocated slots to which the data points will now be assigned to with n/k slots per cluster. The assignment step is then solved using the Hungarian algorithm [13]. After the assignment step, partitioning of the cluster slots in centroid locations in each cluster is done where the centroids are drawn randomly as in original k-means and change in each iterative step, one centroid per cluster. The update step goes in the same way it was traditionally in k-means. The proposed algorithm comes in Balance constrained category.

Constrained k-means[5]: Bradley et al worked on the poor local solutions of the k-means clustering with the resultant clusters being empty or containing very few points, more possibly when n≥10 and k≥20; n being the number of dimensions and k the number of desired clusters. The proposed clustering algorithm aims to pose constraints on the traditional k-means. The constraints force that each cluster should have atleast a predefined number of points in a cluster and hence are k in number for k clusters. The poor solution problems which initially required re-running of the algorithm and re-setting of the empty clusters and again re-running the algorithm will now require less computation time because the clusters will be formed with a sufficient population and shall not remain empty. Guarantee of a predefined population also creates a somewhat balanced cluster. Balanced k-means [4] can be considered as a special case of Constrained k-means with cluster sizes set equal.

Size-Constrained k-means[6]: Zhu and Li proposed that the k-means clustering problem can give better performance if there are imposed size constrains on the algorithm such that each cluster should have a fixed pre-defined size. It uses a prior knowledge of distribution of data to assign the constraints of size. Based on this, the partition of data points should satisfy the given constraint. These constrained clustering problems can be transformed into linear programming optimization by a proposed heuristic procedure.

2.2 *Balanced Driven Algorithms*

Frequency Sensitive Competitive learning (FSCL)[7]**:** Banerjee and Ghosh worked on the limitation of the k-means and spherical k-means clustering in handling high-dimensionality data for competitive learning. Even though spherical k-means aim to normalize the high dimensionality data, the clusters generated fail to show balance when the desired number of clusters is large. For competitive learning, a mixture of von Mises-Fisher distributions was used as the generative model to derive the spherical k-means algorithm. For the applicability to static data and good quality and balance in clustering, three FSCL algorithms were proposed. In brief, the proposal used multiplicative bias for FSCL. In FSCL with additive bias algorithm [8], additive bias was used instead.

Cluster sampled data [9]: Banerjee and Ghosh worked on further scaling the balance constrained clustering algorithms. The proposed algorithm takes three steps to converge. First is sampling of the data points in such a way that the sampled set should be able to represent each of the cluster and with high probability. Second step involves clustering the sampled data which is less than the original data and would hence be efficiently clustered using the proposed soft variant of k –means which satisfies the balancing constraints. The final step populates the clusters formed in step second with the not sampled data in the first step. A less greedy algorithm is proposed for the same which satisfies the balance constraints along with some readjustments and may or may not have any relation with the clustering algorithm used in step second. The data picked for clustering here can be taken irrespective of any domain and the resultant clustering will show outperformed results.

Cuts and fractions: Ratio cut [10] algorithm finds natural partitions and does not require an exact bisection. The ratio cut is defined by the following cost function with P_i as partitions.

$$RatioCut(P_1, \ldots, P_k) = \sum_{i=1}^{k} \frac{cut(P_i, \bar{P}_i)}{|P_i|}$$

The numerator of the function shows minimum-cut criterion with the output showing even partition.

The SR Cut algorithm [11] aims at clustering data with a prior knowledge of the number of clusters expected. The SR-Cut has a cost function summating the inter-cluster similarity and a regularization term measuring the relative sizes of two clusters. A tradeoff between the two terms participating in the cost function then deduces an optimal partitioning which can be further optimized for controlling the size of the generated clusters. The optimization is done through the adjustment of weights of the regularization term and is found to be a NP-Complete problem.

Submodular fractional programming in [12] is a ratio of two submodular function aiming balance as the objective function for regularizing cluster sizes. Ratio cut [10] is its special case. The proposed algorithm is combined with the discrete Newton method and aims to minimize the difference of the two submodular functions involved this objective function. The algorithm provides flexibility in clustering setups by being applied to the objective function which has any number of submodular functions in both numerator and denominator.

3. BALANCING ACCORDING TO PARAMETERS

Parameters involved in a clustering algorithm for balance with related balanced clustering algorithms can be listed as follows

- Size of clusters formed[5][6]
- Number of points in a cluster[6]
- Prior knowledge of data[9][11]
- Modifying objective function[14][11][15]

MinMax k-means [14]: Tzortzis and Likas proposed a balanced k-means clustering algorithm which involves working on the bad initialization problem of the k-means algorithm by altering its objective function. The bad initialization problem is dealt upon by introducing weights to each cluster. The algorithm also modifies the objective of clustering to minimize the maximum intra-cluster variance and not sum of the intra cluster variances. Higher weights are allocated to clusters with larger variance and the intra cluster variance is also computed in a weighted manner. Associated weights help in minimizing clusters with larger variance and the resultant clustering does not involve such clusters. The weights are not user defined and can be learnt by the algorithm itself during iterations of the assignment step of the algorithm. The input provided to the algorithm is a parameter deciding the degree by which penalizing of the clusters having larger variance can be done; penalty being implied through learnt weights.

Fast Balanced k-means [15]: Seeing the issue of the high computational time of the k-means clustering algorithm, authors in [15] aim to eradicate this issue by proposing a Fast-Balanced k-means (FBKmeans). The resultant clustering is achieved in a comparatively low computational time while retaining the other results of the k-means. The proposal involves optimization of the objective function of the k-means algorithm for determining new cluster centres.

4. BALANCING ACCORDING TO APPLICATIONS

Achieving balance in clustering is a prerequisite of most of the applications involving divide and conquer methods with the divide step performed using clustering. Examples of the same are circuit design[10] and photo-query[8]. Balance is desirable in the popular Travelling Salesman problem too where all the salesman should possess equal load of work [16,17]. Wireless sensor networks turn to clustering for prolonging their lifecycle. Seeing that the existing clustering algorithms do no assure of the maximum prolongation of network lifetime, Chawla and Verma in [18] aim to propose a balanced k-means for the issue keeping energy and space equivalent distributions as the basis of the proposal. Balance is also necessary in Logistics planning for improving the service delivery routes. Customer deliveries are successful if the customers are grouped into clusters satisfying conditions of balance and connection. The underlying objective has been worked upon by Cao and Grover in [19] who then proposed two clustering algorithms for the same using special procedures for exploiting Thiessen polygons.

5. CONCLUSION

Achieving balance in clusters formed through the traditional k-means clustering algorithm is the basic objective of this paper. Balance can be achieved through clusters having equal size or importance. This paper sums up the existing research works headed in this direction based on the need for balance, the parameters of the k-means algorithm or the applications where k-means can be used for balance. Each proposed work discussed shows comparatively better results than the conventional k-means clustering algorithm thereby portraying the importance and advantage of having balance in the resultant cluster formation.

REFERENCES

[1] E.W.Forgy, “Cluster analysis of multivariate data: efficiency v/s interpretability of classifications”, *Biometrics*, **21**, 768–769, 1965.

[2] X. Wu et al. “Top 10 algorithms in data mining”, Knowledge and Information Systems, 14(1), pp. 1-37, 2008.

[3] Anil K. Jain, “Data Clustering: 50 Years Beyond K-Means”, *Department of Computer Science & Engineering Michigan State University East Lansing,* Michigan 48824 USA

[4] Mikko I. Malinen and Pasi Fränti, “Balanced *K*-Means for Clustering”, Volume 8621 of the series Lecture Notes in Computer Science pp 32-41 and Proceedings of the Joint IAPR International Workshop, S+SSPR 2014, Joensuu, Finland, 2014.

[5] P. S. Bradley, K. P. Bennett, A. Demiriz, “Constrained k-means clustering”, Tech.rep., MSR-TR-2000-65, Microsoft Research, 2000.

[6] S. Zhu, D. Wang, T. Li, “Data clustering with size constraints”. Knowledge-Based Systems 23(8), pp. 883–889, 2010.

[7] A. Banerjee, B. Ghosh, “Frequency sensitive competitive learning for balanced clustering on high-dimensional hyperspheres”, IEEE Transactions on Neural Net-works 15, 719, 2004.

[8] C. T. Althoff, A. Ulges, A. Dengel, “Balanced clustering for content-based image browsing”, GI-Informatiktage 2011. Gesellschaft f ̈ur Informatik e.V., March 2011.

[9] A. Banerjee, J. Ghosh, “On scaling up balanced clustering algorithms”, Proceedings of the SIAM International Conference on Data Mining, pp. 333–349, 2002.

[10] L. Hagen, A. B. Kahng,, “New spectral methods for ratio cut partitioning and clustering”. IEEE Transactions on Computer-Aided Design 11(9), pp. 1074–1085, 1992.

[11] Y. Chen, Y. Zhang, X. Ji, “Size regularized cut for data clustering”. Advances in Neural Information Processing Systems, 2005.

[12] Y. Kawahara, K. Nagano, Y. Okamoto, “Submodular fractional programming for balanced clustering”. Pattern Recognition Letters 32(2), pp. 235–243, 2011.

[13] R. Burkhard, M. Dell’Amico, S. Martello, “Assignment Problems” (Revised reprint), SIAM, 2012.

[14] G Tzortzis, A. Likas, ”The MinMax k-Means clustering algorithm”, Pattern Recognition, Volume 47, pp 2505–2516, 2014. Elsevier

[15] Adel A. Sewisy, Rasha M. Abd ElAziz, M. H. Marghny, Ahmed I. Taloba, “Fast Efficient Clustering Algorithm for Balanced Data” in (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 5, No. 6, 2014.

[16] Y. Liao, H. Qi, W. Li, “Load-Balanced Clustering Algorithm with Distributed Self-Organization for Wireless Sensor Networks”, IEEE Sensors Journal 13(5), 1498–1506, 2013.

[17] L. Yao, X. Cui, M. Wang, “An energy-balanced clustering routing algorithm for wireless sensor networks”, WRI World Congress, 2009.