International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 04, April 2016
www.ijarse.com
IJARSE
ISSN 2319 - 8354

RECENT DEVELOPMENTS IN ONLINE CLUSTERING

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

In this research Paper I am focusing on The Online clustering has emerged as a new field for research. The limitations of the k-means algorithm of handling a fixed amount of known data is solved through the online settings of the algorithm which effectively clusters the arriving data sequence of unknown nature and size. The online clustering has wide applicability in clustering of real data. In recent years, the management and processing of so-called data streams has become a topic of active research in several fields of computer science such as, e.g., distributed systems, database systems, and data mining. A data stream can roughly be thought of as a transient, continuously increasing sequence of time-stamped data. In this paper, we consider the problem of clustering parallel streams of real-valued data, that is to say, continuously evolving time series. In other words, we are interested in grouping data streams the evolution over time of which is similar in a specific sense. This paper discusses the subject and deduces the open problems for research in the field. A brief survey of recent proposals in this direction aiming at solving the discussed problems is presented in the later part of the paper.

Keywords: Clustering, Data Mining, Data streams, fuzzy sets

I. INTRODUCTION

The study of unsupervised learning is motivated by the raw data produced from various data sources but not yet labeled for any classification task. Clustering, an unsupervised machine learning approach, aims to group or cluster similar data together. Need for clustering arises because in today's era where the data is continuously increasing, analyzing it later for purposes like storing, updation, searching and sorting becomes all the more difficult.

The k-means clustering algorithm [1] is one of the most widely used clustering algorithms, mostly because its robustness and simplicity to effectively cluster the given data. Its ease of implementation makes it the first choice of a data analyst who has to cluster randomly put data into meaningful clusters. The applicability of the k-means algorithm can be seen in almost every field of science and engineering like forecasting, market analysis, image segmentation, image processing, real-time decision making and many more. The algorithm begins taking as input n data points to return as output k clusters. The first step of the algorithm involves

initialization of k cluster centres followed by assigning the (n-k) data points to the k clusters based on the shortest Euclidean distance between them. The algorithm continues till the desired k-clusters are formed or no more data remains to be clustered.

However, the algorithm is bounded by a number of limitations, the major of them being the need of the algorithm to know in prior the set of data points to be clustered and the value of "k" and the random initialization of initial centroids. Despite of all the settings, the algorithm still fails to obtain good performance results in terms of theoretical guarantees. An extension of the algorithm, k-means++ by Arthur and Vassilvitskii[2], as an improvement to the initialization problem of the algorithm, is worth mentioning because of its O(log(k)) approximation.

The description of the algorithm presented above corresponds to the offline setting of the algorithm. The input to the algorithm in this case is fixed, known and the output comes in a single pass of the algorithm. But in cases where the data comes in streams like forecasting, stock prices or the data is too big to be accessed sequentially, the online setting of the algorithm is the best option. The data points arriving as input to the online k-means clustering algorithm one by one can be assigned to any one of the k-clusters or can a form a new single cluster. The online clustering model is shown in Fig. 1.

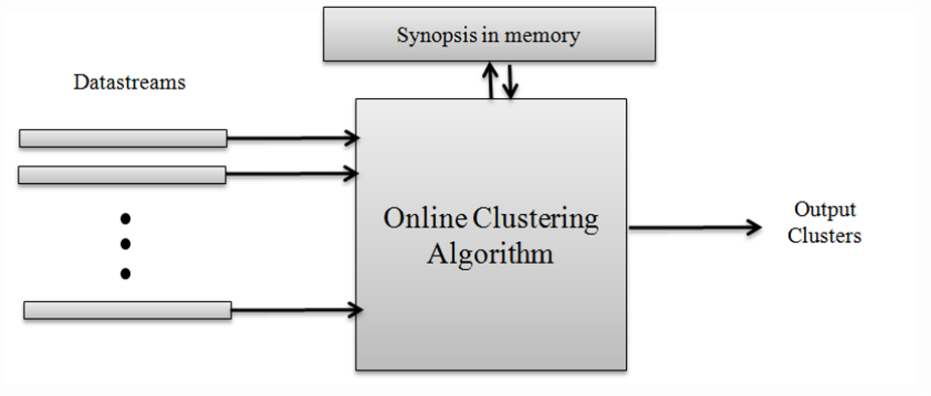

Fig. 1 The online data clustering model

The online setting should however not be confused with the streaming model of the algorithm because of the sequential nature of the data. The online k-means clustering algorithm has a fixed memory and is able to handle an endless stream of data points where each point is visited by the algorithm only once unlike the streaming k-means algorithm where the memory is proportional to the fixed known length of the data stream and the data points can be visited more than once by the algorithm. Testing of the algorithm is done only at the end in the streaming model and at each time step in the online model. Both the algorithms can be converted to each other through processes like divide-and-conquer and random sampling.

This paper deals with exploring the online setting of the k-means algorithm. It is a hot topic of research due to the inability of the traditional and the streaming clustering algorithms to handle arbitrary amount of data that can be consumed in one pass which results in postponing of the clustering decisions. The online model too is subjected to a variety of constraints which is the focus of the recent proposals in this direction. A brief survey discussing the very recent papers is provided in the paper.

II. ABOUT ONLINE CLUSTERING

Online Clustering is a vast field with many open research problems. An online variant of the k-means clustering should focus on eradicating the shortcomings of the original algorithm, should be scalable and allow soft movement of data streams from one cluster to another. The online algorithm should support the objective function of the k-means clustering algorithm and identify "bad" sequences that hamper the growth of the algorithm. Data arriving as input can also not be only stationery and therefore methodologies should be proposed for handling variety of data of the arriving streams.

Based on the discussion above, the various fields of research/ open problems in handling online data streams are listed as

- Approximation of the k-means objective using online learning
- Scalable Clustering of parallel real-valued data streams
- Fuzzy variant of the k-means online clustering algorithm
- Working on limitations of the traditional k-means clustering before jumping to its online versions
- Identifying "bad" sequences
- Applicability of online clustering in various fields
- Handling non-stationery data
- Integration with stream clustering or conversion of both the algorithms into each other

The coming section discusses some of the noteworthy research works proposed recently focus of which is one of the above mentioned open problems.

III. WORKING MODULE OF ONLINE CLUSTERING

Beringer and Hullermeier [3] proposed an online algorithm for clustering incoming parallel data streams where each stream consists of real valued data. The incoming data has to be analyzed along with taking into account, that a delay not more than a specific time is encountered in the process. The proposed online variant of the k-means algorithm is scalable as a result of the online transformation done to the original data. The online transformation is a preprocessing step that calculates the distance between data streams using approximations from original data thereby making the algorithm feasible to be used in a real world environment for clustering of thousand data streams. The authors also proposed a fuzzy variant of the online k-means algorithm to allow shifting of data stream from one cluster to the other smoothly.

Barbakh and Fyfe [4] address the limitations of the traditional k-means algorithm of converging to a local optima and its sensitivity to the initial centroids. The reason they point out is the performance function of the algorithm for which they propose an improvement in the algorithm. The proposed improvement makes the algorithm able to effectively work even in the worst case scenario when all the points are initialized in the same positions. Their idea is based on the knowledge of a data point about the locations of all the remaining data points and its relative positions from the others before it changes its location. This can effectively improve the performance of the algorithm as no clusters will be left free or overcrowded at the convergence of the algorithm. The authors develop a family of new algorithms using the proposed improvement in both online and offline settings. Work in the direction of visualization and topology-preserving mappings is also proposed.

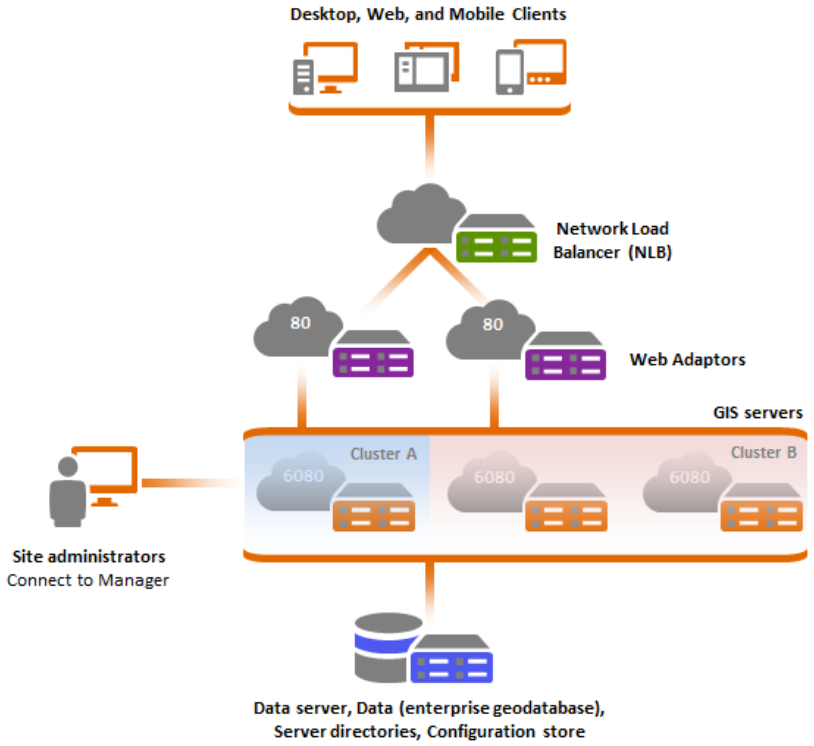

Figure. 2 The Online Clustering Model

Choromanska and Monteleoni [5] worked on online clustering using experts. The aim of the proposal is to compute approximation guarantees of the online clustering algorithm in terms of the k-means objective function. In cases where it needs to be checked which of the batch clustering algorithms are effective for working in online setting, the proposal helps determine the needed effectiveness. The batch clustering algorithms of [6] and [7] are taken as experts and their approximations are calculated based on the computation of approximation to the current value of the k-means objective through each expert. The guarantee is in terms of cost of the entire data stream as the k-means objective desires, even though the setting of the algorithm is now online. The inspiration behind the proposal is the work of Dasgupta based on regret analyses [8]. Another work on regret analyses is done by Gentile et al [9] for content recommendation using clustering of exploration-exploitation strategies.

Khaleghi et al's work [10] in the direction of online time-series clustering is motivated by the fact that the data arriving in sequences, either new or merged with the older ones, come from various distributions and proper distinction of the source of the sequence, if not done by the clustering algorithm, can hamper the performance of the algorithm. Inability of the algorithm to distinguish can be a result of the "bad" sequences with not enough information regarding their generating distributions. In any way, not detecting "bad" sequences can affect the clustering algorithm even if the remaining sequences have proper information or are "good". For the same, a non-parametric asymptotically consistent online clustering algorithm has been proposed that works well for sequences arriving from stationary or ergodic. The algorithm is robust to "bad" sequences with zero error rates and is applicable to real data.

Angie King [11] proposed clustering non-stationery data in online settings. The author notes down the differences non-stationery data clustering makes and how the cost objective function of the k-means algorithm is

not effective for the problem. An online non-stationery data clustering algorithm is proposed in the paper with proved performance and practical guarantees. Liberty et al [12] proposed working in an integrated environment of online clustering model and stream clustering model.

IV. CONCLUSION

The paper provides a brief survey on the recent trends in the field of online clustering. An introduction to the subject pointing out the various open areas of research in the field has been provided in the paper. The survey lists some of the noteworthy research works aiming to further expand and improve the online k-means clustering algorithm.

REFERENCES

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

[2] D. Arthur and S. Vassilvitskii, "k-means++: The Advantages of Careful Seeding", Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, pp. 1027-1035, 2007.

[3] J. Beringer and E. Hullermeier, "Online Clustering of Parallel Data Streams", Data and Knowledge Engineering, Volume 58, no.2, pp.180-204, 2006.

[4] W. Barbakh and C. Fyfe, "Online Clustering algorithms", International Journal of Neural Systems, Volume 18, no. 03, pp. 185-194, 2008.

[5] A. Choromanska and C. Monteleoni, "Online Clustering with experts", Proceedings of the 15th International Conference on Artificial Intelligence and Statistics(AISTATS), pp. 227-235, 2012.

[6] M. Herbster and M.K. Warmuth, "Tracking the best expert", Machine Learning, Volume 32, pp. 151-178, 1998.

[7] C. Monteleoni and T. Jaakkola, "Online Learning of non-stationery sequences", NIPS, 2003.

[8] Sanjoy Dasgupta, "Lecture 6: Clustering in an online streaming setting", Course notes, CSE 291: Topics in unsupervised learning, Section 6.2.3, in http://www-cse.ucsd.edu / ~dasgupta/291/lec6.pdf, University of California, San Diego, Spring Quarter, 2008.

[9] C. Gentile, S. Li and G. Zappella, "Online Clustering of Bandits", 2014. Available at arXiv preprint arXiv: 1401.8257.

[10] A. Khaleghi, D.Ryabko, Jeremie Mary and Philippe Preux, "Online Clustering of Processes", Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 227-235, 2012.

[11] A. King, "Online k-Means Clustering of Nonstationary Data", 15.097 Prediction Project Report, Machine Learning and Statistics, 2012.

[12] E. Liberty, R. Sriharshay and M. Sviridenkoz, "An Algorithm for Online K-Means Clustering", 2015. Available at arXiv:1412.5721v2 [cs.DS]