IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727
PP 77-79
www.iosrjournals.org

Survey On Clustering Method For Randomized Dimensionality Reduction

Miss.Shilpa R. Mokashe[1], Prof. Sanjay B. Thakare[2]
[1] *(Computer Department, RSCOE,Pune/ savitribai phule pune university, India)*
[2] *(Computer Department, RSCOE,Pune/ savitribai phule pune university, India)*

Abstract: *They study the topic of dimensionality reduction for k-means clustering. Dimensionality reduction covers the combination of two approaches first feature selection and second feature extraction. A feature selection-based algorithm for k-means clustering chooses a small subset of the input features and then applies k-means clustering on the selected features. A feature extraction-based algorithm for k-means clustering generates a small set of new artificial features and then applies k-means clustering on the constructed features. Even though the importance of k-means clustering as well as the wealth of heuristic methods addressing it, provably accurate feature selection methods for k-means clustering are not known. On the other hand, two provably accurate feature extraction methods for k-means clustering are identified in the literature that is one is based on random projections and the other is based on the singular value decomposition (SVD). This paper makes further advancement in the direction of a better understanding of dimensionality reduction for k-means clustering. Namely, they present the first provably accurate feature selection method for k-means clustering and, in addition, they present two feature extraction methods. The first feature extraction method is based on random projections and it improves upon the previous results in terms of time complexity and number of features needed to be extracted. The second feature extraction method is depends upon fast approximate SVD factorizations and it also improves upon the existing results in terms of time complexity. The proposed algorithms are randomized and present constant-factor approximation guarantees with respect to the optimal k-means objective value.*
Keywords: *clustering, dimensionality reduction, randomized algorithms.*

I. Introduction

Clustering is ubiquitous in science and engineering with numerous application domains ranging from bioinformatics and medicine to the social sciences and the web Perhaps the most well-known clustering algorithm is called as "*k*-means" algorithm or the Lloyd's method. Lloyd's method is an iterative expectation-maximization type approach that attempts to deal with the following objective. given a set of Euclidean points and a positive integer *k* corresponding to the number of clusters, split the points into *k* clusters so that the total sum of the squared Euclidean distances of each point to its nearest cluster centre is minimized Due to this intuitive objective as well as its *effectiveness*, the Lloyd's method for *k*-means clustering has become enormously popular in applications. As of late, the high dimensionality of present day huge datasets has given a significant test to the outline of effective algorithmic answers for k-means bunching. To begin with, ultra-high dimensional information power existing calculations for k-means grouping to be computationally wasteful, and second, the presence of numerous immaterial components may not permit the ID of the significant fundamental structure in the information . Specialists have tended to these snags by presenting component choice and highlight extraction procedures. Highlight determination chooses a (little) subset of the real components of the information, though include extraction develops a (little) set of counterfeit components taking into account the first elements.

II. Related Work

In this paper [1], author outlined, Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm Bunching investigation strategy is one of the primary expository routines in information mining, the system for grouping calculation will impact the bunching results specifically. This paper talks about the standard k-means bunching calculation and examines the deficiencies of standard k-implies calculation, for example, the k-means grouping calculation needs to compute the separation between every information article and every single group focus in every cycle, which makes the productivity of bunching is not high. This paper proposes an enhanced k-implies calculation keeping in mind the end goal to explain this inquiry, requiring a straightforward information structure to store some data in each cycle, which is to be utilized as a part of the following integration. The enhanced technique abstains from figuring the separation of every information article to the group focuses repeal, sparing the running time. Test results demonstrate that the enhanced strategy can

adequately enhance the rate of grouping and exactness, diminishing the computational unpredictability of the k-implies.

In this paper[2], Top 10 algorithms in data mining they presents the top 10 data mining algorithms identified by the IEEE International Conference on Data Mining (ICDM) in December 2006: C4.5, *k*-Means, SVM, Apriority, EM, Page Rank, Gadabouts, *k*NN, Naive Bayes, and CART. These top 10 algorithms are among the most influential data mining algorithms in the research community. With each algorithm, they provide a description of the algorithm, discuss the impact of the algorithm, and review current and further research on the algorithm. These 10 algorithms cover classification.

In this paper [3], They demonstrate the presence of little corsets for the issues of figuring k-middle and k-means bunching for focuses in low measurement. At the end of the day, they appear that given a point set P in IRd, one can figure a weighted set S P, of size O(k"−d log n), such that one can figure the k-middle/means bunching on S of on P, and get a (1 + ")- estimation. Subsequently, they enhance the speediest known calculations for (1 + ") - estimated kmeans furthermore, k-middle. Our calculations have straight running time for a settled k and ". In expansion, it is possible to keep up the (1 + ")- surmised k-middle or k-means grouping of a stream when focuses are by and large just embedded, utilizing polylogarithmic space and redesign.

In this paper[4], They show that there exists a (k, ")-coreset for k-median and k-means clustering of n points in IRd, which is of size independent of n. In particular, they construct (k, ")- coreset of size.

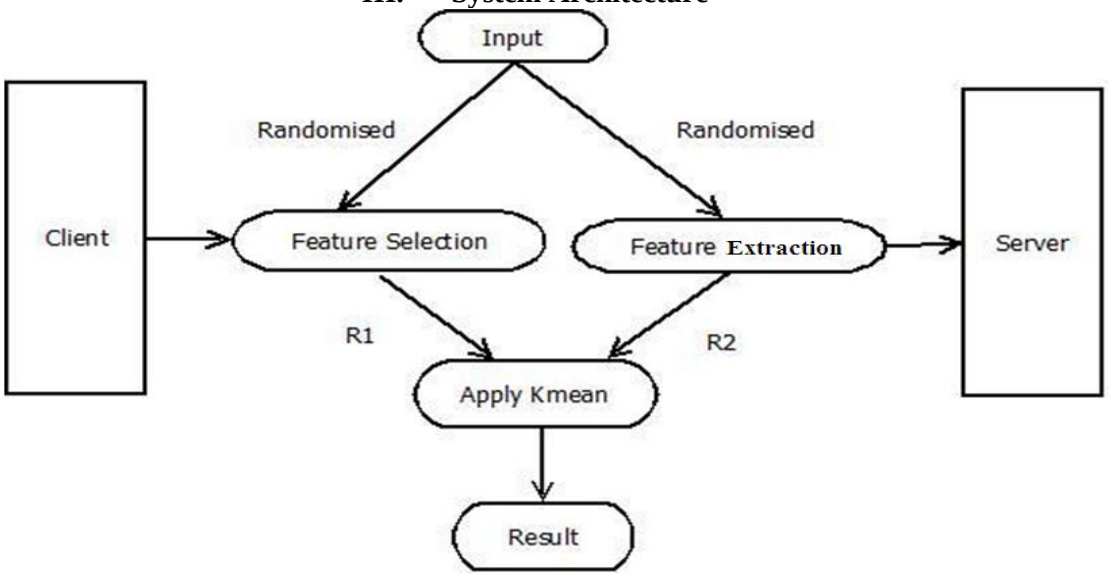
In this paper[5], author studied , they portray a straightforward irregular inspecting based method for creating inadequate grid approximations. Our technique and examination are to a great degree basic: the investigation uses nothing more than the Chernoff-Hoeffding limits. In spite of the straightforwardness, the guess is equivalent also, some of the time superior to anything past work. Our calculation registers the scanty framework estimation in a solitary ignore the information. Further, the greater part of the sections in the yield framework are quantized, and can be concisely spoken to by a bit vector, in this manner prompting much reserve funds in space.

In this paper[6], the Johnson-Lindenstrauss irregular projection lemma gives a basic approach to lessen the dimensionality of an arrangement of focuses while around protecting their pair wise separations. The most direct utilization of the lemma applies to a limited arrangement of focuses, however late work has developed the procedure to relative subspaces, bends, and general smooth manifolds. Here the instance of irregular projection of smooth manifolds is considered, furthermore, a past investigation is honed, lessening the reliance on such properties as the complex's greatest ebb and flow.

In this paper [7], Inspired by applications in which the information may be planned as a network, they consider calculations for a few normal straight variable based math issues. These calculations make more efficient utilization of computational assets, for example, the calculation time, arbitrary access memory (RAM), what's more, the quantity of disregards the information, than do already known calculations for these issues. In this paper, author devise two calculations for the framework duplication issue. Assume An and B (which are m × n and n × p, separately) are the two info networks. In our fundamental calculation, they perform c free trials, where in every trial they arbitrarily test a component of {1, 2,...,n} with a suitable likelihood circulation P on {1, 2,...,n}. They shape a m × c lattice C comprising of the examined sections of An, each scaled fittingly, and we shape a c × n lattice R utilizing the relating columns of B, again scaled properly. The decision of P and the section and column scaling are pivotal elements of the calculation. At the point when these are picked reasonably, they demonstrate that CR is a decent estimate to AB. All the more absolutely, they demonstrate.

In this paper [8], author proposed that this, they distinguish two issues included in adding to a mechanized component subset determination calculation for unlabeled information: the requirement for discovering the quantity of groups in conjunction with highlight choice, and the requirement for normalizing the inclination of highlight determination criteria with deference to measurement. They investigate the component choice issue and these issues through FSSEM (Feature Subset Selection utilizing Expectation-Maximization (EM) bunching) and through two distinctive execution criteria for assessing hopeful element subsets: scramble detachability and most extreme probability. They present verifications on the dimensionality predispositions of these element criteria, and present a cross-projection standardization conspire that can be connected to any foundation to enhance these predispositions. Our examinations demonstrate the requirement for highlight choice, the requirement for tending to these two issues, what's more, the adequacy of our proposed arrangements

III. System Architecture



IV. Conclusion

They considered the issue of dimensionality decrease for k-means bunching. The greater part of the current results in this theme comprise of heuristic methodologies, whose magnificent observational execution cannot be clarified with a thorough hypothetical investigation. In this paper, our emphasis was on dimensionality lessening systems that function admirably in principle. They displayed three such methodologies, one component choice technique for k-implies what's more, two element extraction techniques. The hypothetical investigation of the proposed systems depends on the way that dimensionality diminishment for k-means has profound associations with low-rank approximations to the information framework that contains the focuses one needs to group. Author clarified those associations in the message and utilized cutting edge quick calculations to process such low rank approximations and outlined quick calculations for dimensionality diminishment in k-implies.

References

[1]. J. A. Hartigan, *Clustering Algorithms*. New York, NY, USA: Wiley, 1975.
[2]. X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, 2008.
[3]. A. Kumar, Y. Sabharwal, and S. Sen, "A simple linear time (1 + _)- approximation algorithm for *k*-means clustering in any dimensions," in *Proc. 45th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, 2004,pp. 454–462.
[4]. S. Har-Peled and A. Kushal, "Smaller coresets for *k*-median and *k*-means clustering," in *Proc. 21st Annu. Symp. Comput. Geometry (SoCG)*, 2005, pp. 126–134.
[5]. S. Arora, E. Hazan, and S. Kale, "A fast random sampling algorithm for sparsifying matrices," in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques* (Lecture Notes in Computer Science), vol. 4110. Berlin, Germany: Springer-Verlag, 2006, pp. 272–279. [Online]. Available: http://dx.doi.org/10.1007/11830924_26.
[6]. K. L. Clarkson, "Tighter bounds for random projections of manifolds," in *Proc. 24th Annu. Symp. Comput. Geometry (SoCG)*, 2008, pp. 39–48.
[7]. P. Drineas, R. Kannan, and M. Mahoney, "Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication," *SIAM J. Comput.*, vol. 36, no. 1, pp. 132–157, 2006.
[8]. C. Boutsidis, M. W. Mahoney, and P. Drineas, "Unsupervised feature selection for the *k*-means clustering problem," in *Neural Information Processing Systems*. Red Hook, NY, USA: Curran & Associates Inc., 2009.
[9]. STUART P. LLOYD, `` Least Squares Quantization in PCM", IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. IT-28, NO. 2, MARCH 1982.
[10]. Sariel Har-Peled, Soham Mazumdar, `` Coresets for k-Means and k-Median Clustering and their Applications".
[11]. Isabelle Guyon, Steve Gunn, Asa Ben Hur and Gideon Dror, `` Result Analysis of the NIPS 2003 Feature Selection Challenge".
[12]. P. Drineas, Alan Frieze, Ravi Kannant, Santosh Vempalas, V. Vinayr, `` Clustering in large graphs and matrices".