ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 6, Issue 2, February 2021

# An Extensive Study on Machine Learning Techniques for Large Scale High Dimensionality Data Clustering

[1] C.Deepa
[1] Assistant Professor, PSG College of Arts & Science, Coimbatore-14, Tamil Nadu, India

*Abstract:* - **The rapid growth of electronic documents with large scale and high dimensionality is challenging task as data is unstructured and it required more time and much effort to cluster those documents in many domains. Many clustering algorithms using machine learning algorithm have been developed to address those documents with a very large sample size or with a very high number of dimensions, but they are often impractical and great challenge when the data is large in both aspects and leads to curse of dimensionality, data noise, data sparsity and data scalability issues as it effects the effectiveness and efficiency. Data transformation using Heuristic and hybrid technique have proposed to handle categorical and numeric attributes simultaneously, and scales well with the dimensionality and the size of data on distance between data points. In this paper, an extensive study on machine learning techniques on employing hybrid and ensemble model to handle large scale high dimensional data on aspects of data pre-processing, dimensionality reduction, feature selection and feature extraction and finally clustering has been estimated in detail. The clustering algorithm majorly classified as partition-based clustering, Kernel based clustering hierarchical based clustering, Density based clustering and subspace clustering. These analyses provide the solution for fast data-space reduction and an intelligent sampling to cluster the data effectively on various objective functions and optimal solutions configurations to alleviate mentioned issues. Experimental analysis on machine learning based data clustering model on multiple setting has been carried out on the various data sets using performance metric such as Euclidean distance, accuracy, execution time and silhouette index.**

**Key Words: Machine Learning, Text Clustering, Large Scale Data, High Dimensional data, Curse of dimensionality**

## INTRODUCTION

Data clustering [1] is an unsupervised process of grouping and portioning elements into several subset of similar object together on exploratory data analysis. Clustering can be considered as the most important unsupervised learning problem. Clustering deals with finding a primitive structure in a collection of unlabelled data. The objective of the data clustering technique is to determine the intrinsic grouping of the large scale unlabelled data. The similarity between data objects of the dataset can be measured with the computation of distance values. Specifying the distance measures for the high dimensional data is becoming very trivial because it holds different data values in their corresponding attributes.

Cluster analysis is to discover useful patterns by explaining and characterizing the partition. Initially the high dimensionality and large scale of the data presents a specific challenge for clustering algorithm on basis of curse of dimensionality and data sparsity especially as it is based on Euclidean distance. Secondly data is usually noisy, containing some irrelevant or redundant information which hides the cluster patterns on dynamic changing of the data distributions leads to empty space phenomena and concentration of distance. Finally many machine learning algorithms that rely on the traditional distance measure are sensitive to different units of the attributes [3]. The distance based measures on high dimensional data sets to be sparse and distributions grows exponentially on the number of dimensions. To work efficiently with large data sets, the algorithms must have high scalability.

Clustering is useful in several exploratory pattern analysis, grouping, decision making and machine learning situations including data mining, document retrieval, image segmentation and pattern classification. High dimensional data clustering are abundant in processing using machine learning algorithm. Exploring various data clustering algorithm based on supervised and unsupervised learning techniques has envisioned learning the relation between performances of the cluster result. In this paper, different types of data clustering model for high dimensional data has been analysed and evaluated in detail along data pre-processing, feature reduction, feature extraction and selection, clustering and finally distance measures. These models prove the quality of the cluster and homogeneity.

The rest of paper is structured as follows, section 2 describes the definition of the data mining process, where section 3 describes the types of the data clustering model for high dimension data, while section 4 presents the review of literature on data clustering technique on real time applications with its advantages and limitation has been analysed in depth. Finally Section 5 concludes the work.

## 2. Definition

In this section, definition of the various data mining process has been provided as those considered as preliminary step of the clustering of high dimensional data with irrelevant data in the dataset. They are

### 2.1 Data pre-processing

Data pre-processing is one of the important preliminary step in machine learning towards clustering the data which aim to reduce and normalize the data. It deals with data preparation and data transformation. It involves cleaning, integration, transformation and reduction. Feature reduction is employed to remove the outlier data. In addition, it missing data prediction has been also carried out for the categorical as well for numerical data[3].

### 2.2 Feature Selection

Feature selection is one of the most used techniques to reduce dimensionality and aims to choose a small subset of the relevant features from the original ones according to certain relevance evaluation criterion[4]. A feature selection method consists of four basic steps, namely, subset generation on the candidate features based on strategies, subset evaluation based on criterion, stopping criterion, and result validation. Feature weighting was applied as feature selection in tasks where features vary in their relevance score.

## 3. Types of Data Clustering model

In this section, different category of data clustering model has been analysed on the attributes type of the data on high dimensional dataset. They are as follows

### 3.1. Partitioned based Clustering

Partitioned based clustering partition the n objects into a set of k non-overlapping groups. Partitioned clustering algorithms use an iterative approach to group the data into a k number of clusters by minimizing the objective function[5]**.** It process as finding the nearest neighbour to all the points by computing the distance between points and assign it to nearest seed point. Category of the clustering is K means, PAM, CLARA and Fuzzy C Means.

### 3.2. Hierarchical based Clustering

Hierarchical clustering algorithms have been employed to nested clusters. It operates in two mode which is agglomerative mode and division mode[6]. Agglomerative mode is a bottom up method of clustering which start with single data point as its own cluster and merging the most similar pair of clusters successively till a final cluster is obtained that has all the data point. Division mode is top down clustering method which starts with all data points contained as one cluster and recursively dividing each cluster into small cluster.

### 3.3. Density based Clustering

Density based clustering is to discover clusters of arbitrary shape on condition that each cluster contains the density points and density of inner cluster should be higher than outside the cluster as the condition[7]. The key idea of density-based clustering is that the numbers of data objects in the "neighborhood" are considered to determine density. The category of the clustering is DBSCAN. It works only on the locality of the point assumes that the objects inside of the clusters are randomly distributed and it does not need any input parameter.

### 3.4. Subspace Clustering

Subspace clustering methods is employed to search for clusters in a particular projection of the data on ignore irrelevant attributes[8]. Subspace clustering approaches integrates feature selection into the clustering process. An approach might be to search through all possible subspaces and use cluster validation techniques to determine the subspaces with the best clusters. Subspace clustering must evaluate features on only a subset of the data, representing a cluster

## 4. Review of literature

In this section, various literatures employed for clustering using high dimensional data has been examined in detail.

- Nenad Tomasev[9] et.al proposed a model that uses the role of hubness for high dimensional data clustering towards distinguishing distances between data points and observes a lower dimensional feature subspace. It contains points (hubs) that frequently occur in k-nearest- neighbor lists of other points, can be successfully exploited in clustering. It can be used effectively as cluster prototypes or as guides during the search for centroids-based cluster configurations.
- Sangdi[10] Lin et.al proposed a model that uses the CRAFTER which is able to handle categorical and

numeric attributes simultaneously, and scales well with the dimensionality and the size of datasets. The concept of the class probability estimates is utilized to identify the representative data points for clustering. It is sensitive to irrelevant attributes, noise and outliers, and it scales well with the size of datasets.

- Punit Rathore[11] et.al proposed a model that uses the Rapid Hybrid Clustering Algorithm which is employed for clustering large volumes of high-dimensional data. It is to compare the cluster distributions in samples obtained from three sampling strategies: random sampling, MMRS sampling in the p dimensional up space, and MMRS sampling in the q dimensional downspace.
- C.C. Aggarwal[12] et.al proposed a model that uses the generalized projected clustering technique may also be viewed as a way of trying to redefine the clustering for high dimensional applications by searching for hidden subspaces with clusters which are created by inter-attribute correlations.
- Kaban[13] proposed model which employs Non-Parametric Detection of Meaningless Distances in High Dimensional Data which uses phenomena of high dimensional data processing, analysis, retrieval, and indexing, which all rely on some notion of distance or dissimilarity. It uses finite-dimensional characterisation of the distance concentration phenomenon that recovers previous results in the limit of infinite dimensions.

## 5. Tabular view of the review of literatures

| SI.No | Problem | Objective | Technique | Advantages |
|---|---|---|---|---|
| 1 | Curse of dimensionality | Need to compute lower dimensional feature space to cluster | Hubness clustering | Model is easy to interpret.<br>Fast and efficient computation.<br>Not affected by irrelevant features |
| 2 | Class probability issue | Need to identify representative data point for clustering. | Tree based Ensemble Model | Data dimensionality has been reduced dependency. |
| 3 | Cluster distribution issues | It uses the sampling strategies on top down approach. | Rapid Hybrid Clustering using DBSCAN | Capable of learning any rare relationship on the feature instances. |
| 4 | Hidden Space | Redefine Clustering using inter attribute correlation | Projected Clustering | Interrelation is high |
| 5 | Distance of dissimilarity in feature space | Finite dimensional characterization | Non-Parametric Detection | Cluster distance is low and highly scalable. |

## 6. Conclusion

An extensive study on machine learning algorithm for data clustering has been carried out on high dimensional data and those approaches towards clustering has been carried out in detail in this work. On analysis, it came to conclusion that, high dimensional data has been clustered with high accuracy and produces high scalability. The curse of dimensionality and sparsity has been eliminated on inclusion of cluster constraints. Finally experimental analysis of analysed technique to cluster of the high dimensional data using machine learning model has proved on different strategies to demonstrate the effectiveness and robustness of the approaches.

## References

[1] L. Parsons, E. Haque, and H. Liu, “Subspace clustering for high dimensional data: a review,” ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 90–105, 2004.

[2] T. Shi and S. Horvath, “Unsupervised learning with random forest predictors,” Journal of Computational and Graphical Statistics, vol. 15, no. 1, 2006.

[3] B. Azarnoush, J. M. Bekki, G. C. Runger, B. L.

Bernstein, and R. K. Atkinson, "Toward a framework for learner segmentation," Journal of Educational Data Mining, vol. 5, no. 2, pp. 102–126, 2013.

[4] Q. Zhang and I. Couloigner, "A new and efficient k-medoid algorithm for spatial clustering," Computational Science and Its Applications–ICCSA 2005, pp. 207–224, 2005.

[5] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for kmedoids clustering," Expert systems with applications, vol. 36, no. 2, pp. 3336–3341, 2009.

[6] J. Ji,W. Pang, C. Zhou, X. Han, and Z.Wang, "A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data," Knowledge-Based Systems, vol. 30, pp. 129–135, 2012.

[7] Z. He, X. Xu, and S. Deng, "Attribute value weighting in k-modes clustering," Expert Systems with Applications, vol. 38, pp. 15 365– 15 369, 2011.

[8] N. Tomasev, M. Radovanovic, D. Mladenic, and M. Ivanovic, "The role of hubness in clustering high-dimensional data," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 3, pp. 739–751, 2014

[9] Sangdi Lin, Bahareh Azarnoush, George C. Runger"CRAFTER: a Tree-ensemble Clustering Algorithm for Static Datasets with Mixed

[10] Attributes and High Dimensionality"IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL.5 , NO. 8, 2017

[11] Punit Rathore and Dheeraj Kumar" A Rapid Hybrid Clustering Algorithm for Large Volumes of High Dimensional Data "IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, vol.12, issue.14, 2019

[12] C.C. Aggarwal and P.S. Yu, "Finding Generalized Projected Clusters in High Dimensional Spaces," Proc. 26th ACM SIGMOD Int'l Conf. Management of Data, pp. 70-81, 2000.

[13] Kaban, "Non-Parametric Detection of Meaningless Distances in High Dimensional Data," Statistics and Computing, vol. 22, no. 2, pp. 375-385, 2012.