Data Clustering & High Dimensionality – A Review

Ms. M Udaya Rani
Associate Professor
Department of IT, SVIT
Secunderabad, India

Dr. P V Kumar
Professor
Department of CSE, OUCE
Hyderabad, India

Dr. S Durga Bhavani
Professor
School of IT, JNTU-H
Hyderabad, India

Abstract— *The huge accumulations of data in traditional databases and the application of Data Analysis has led to Data Mining. With the wild usage of the Internet, the shape and size of data changed drastically. Data is no longer low dimensional. High dimensionality defeats all the works done on low dimensional data. This paper looks at the various reviews, surveys done on Data Clustering and High Dimensionality and also looks at the current trends.*

Keywords— Data Mining, Clustering, High Dimensional Data

I. INTRODUCTION

Data mining is an evolving research field, and is diversifying to many other related disciplines. Data has been studied and worked with in various ways. Traditional databases concentrated on storage and management of Data. Data Mining typically studies the data repositories with Data Analysis perspective and arrives at -

- Generalization/Characterization - The data is mined to figure out the general description of the data.
- Discrimination - Class discrimination involves the discrimination of characterization of target and contrast class.
- Association - Association analysis involves figuring out frequent patterns and framing of association rules based on the frequent patterns derived.
- Classification - Classification will identify the class information of previously unseen data by applying the learned knowledge of previously classified data. It is a supervised technique.
- Clustering - Clustering identifies similar data objects and groups them into clusters using unsupervised techniques.
- Outlier Detection - Outlier analysis will figure out deviations in data.

II. CLUSTERING

A. Data Clustering

Data Clustering is one of the oft used data mining functionality. It tries to group similar data into clusters with no prior knowledge about any class information and hence a totally unsupervised technique. Many techniques exist by which we can cluster data, like

- Distance based – k-means, k-medoids, k-medians algorithms
- Hierarchical – linkage clustering, AGNES, DIANA algorithms
- Density based – DBSCAN, OPTICS algorithms
- Grid based – STING algorithm
- Probabilistic model based – Expectation-maximization algorithm

And many more

Most of the classical clustering techniques look at the data traditionally – numerical, categorical and might not be suitable for mixed data. They might also have a predefined notion of the number of clusters to derive. None of these can automatically label the clusters. These are best suited for low dimension data i.e., may be up to 10 dimensions. The present day data is typically not traditional and is of High Dimension in nature i.e., typically have more than 10 dimensions and may go up to thousands of dimensions.

B. High Dimensional Data Clustering

The properties of high dimensionality are poorly understood and clustering high dimensional data is challenging. High dimensional data is subjected to the curse of dimensionality. In very simple terms if a data objects are represented by a single dimension, they can be plotted on a line, two dimensions a x-y plane, three dimensions a x-y-z plane, but as the dimensions increase the space occupied by the data objects grows explosively. The data objects are lost in the multi-dimensional space, i.e., become sparse because of the increase in the volume of the space. The distance functions like Euclidean distance becomes meaningless. The similarity between the objects is lost and only when we look into subspaces can we find similar objects. Because of the large number of dimensions, some of the dimensions may be irrelevant and correlated. Similarities may be seen among the correlated dimensions, resulting in misleading correlations. High dimension visualization is not intuitive and is challenging. The complexity of

the algorithms becomes exponential, and inapplicable for real applications. And to top it the computation cost of processing is way too expensive.

Some of the popular approaches used to cluster High dimensional data are:

- Dimensional reduction - feature selection (FAST), feature extraction (PCA)
- Correlation based - 4C, COPAC
- Subspace clustering – CLIQUE algorithm
- Projected clustering – PROCLUS algorithm
- Hybrid clustering – FIRES algorithm

III. Earlier Works

Some of the earlier work done on High-Dimensional Data Clustering:

- The Curses and Blessings of Dimensionality [1] -

This paper summarizes High Dimensional Data Analysis. When we have traditional data from traditional data systems, it is typically low dimension in nature. For this data we can assume that $D < N$, and $N \rightarrow \infty$ and data analysis can get some exact distributional results. When we have non-traditional complex data it is high dimensional data and the data analysis done for low dimensional data no longer holds good and fails as $D > N$. Even worse, they envision an asymptotic situation in which $N \rightarrow \infty$ with D fixed, and that also seems contradicted by reality, where we might even have D tending to ∞ with N remaining fixed.

- Challenges of High Dimensional Clustering [2] –

This paper summarizes the challenges and the techniques available to cluster High dimensional data.

- High Dimension Data Clustering [3] –

This survey looks into subspace clustering and clarifies on the different problem definitions in general; the difficulties encountered in research; the assumptions, heuristics, and intuitions of different approaches; and the solutions to the problems.

- Survey Of Big Data Clustering Algorithms [4] –

A big volume of data or big data has its own deficiencies as well. They need big storages and this volume makes operations such as analytical operations, process operations, retrieval operations, very difficult and hugely time consuming. This paper discusses these problems.

IV. Recent Works

Present day data clustering focus on applications of the techniques and there are many works done in this direction.

- Discovering Overlapping Communities by Clustering Local Link Structures [5] –

The existing community detection methods commonly face two challenges: incorrect base-structures and incorrect membership of weak-ties. To overcome both problems, a Local link structure (LLS) clustering based method for overlapping community detection is proposed.

- A survey on online Stock forum using subspace clustering [6] –

Financial stock Data Analysis and future prediction in terms of Sentiments is great challenge in the big data research. For unlabeled opinion, opinion classification in terms of unsupervised learning algorithm will lead to classification error as data is sparse and high dimensional. To overcome this problem, the sentiment analysis
to extract the opinion of each word in the stock data is proposed. The singular value decomposition is used to resolve the inconsistent constraints correlating to the large dimensions, and dimensionally reduced feature set is been used. The dimensionally reduced feature set is classified into clusters through employment of Principle component analysis with utilization of the domain knowledge.

- Implementation of scalable K-Means++ clustering for passengers temporal pattern analysis in public transportation system (BRT Trans Jogja case study) [7] –

Smart Card Automated Fare Collection System (SCAFCS) which is currently used as e-ticketing in Trans Jogja public transport is used to analyze passengers pattern with data mining approaches. This paper applied SCAFCS data preprocessing with data warehouse mechanism and implemented Hadoop Platform as distributed computing to improve K-Means++ clustering performance on large datasets scalability; in this case, SCAFCS Trans Jogja has a large dataset (volume) and rapid growth data (velocity). The clusters are used to analyze passengers pattern based on the dimensions of time (temporal), segmentation of passengers (structure) to determine the variability of passengers based on the card they used and transaction peak on boarding location (spatio).

- An Improved K-means Algorithm for Document Clustering [8] –

K-Means algorithm has a major shortcoming of high dimensional and sparse data. This paper proposes a K-Means algorithm based on Sim Hash. After preprocessing, Sim Hash is used to calculate the feature vectors extracted, and then the fingerprint of each text is obtained. Sim Hash not only reduces the dimension of the text, but also directly calculates the Hamming distance between the fingerprints as the vector distance and based on the Hamming distance, it determines which cluster the data is belongs to.

- An extractive text summarization technique for Bengali document(s) using K-means clustering algorithm [9] –

This paper presents a method for text summarization
which extracts important sentences from a single or multiple Bengali documents. The input document(s) should be pre-processed by tokenization, stemming operation etc. Then, word score is calculated by Term-Frequency/Inverse Document Frequency (TF/IDF) and sentence score is determined by summing up its constituent words' scores with its position. For single or multiple documents, K-means clustering algorithm has been applied to produce the final summary. The experimental result shows satisfactory outputs in comparison to the existing approaches possessing linear run time complexity.

- Clustering Smart Card Data for Urban Mobility Analysis [10] –

In this paper, two approaches to cluster smart card data, which can be used to extract mobility patterns in a public transportation system are proposed. Two complementary standpoints are considered: a station-oriented operational point of view and a passenger-focused one. By applying the approaches to a real data set, it illustrates how they can help reveal valuable insights about urban mobility.

- Clustering for Similar Recipes in User-Generated Recipe Sites Based on Main Ingredients and Main Seasoning [11] –

This paper, proposes a clustering method for user-generated recipe sites based on page structure and main ingredient and main seasoning of the food. It provides a means of classifying the user search results according to similar pages. The experiment conducted to measure the benefits of the proposed method presents the results of benefits of the method, which classifies similar recipes based on the main ingredients and main seasonings.

- High-dimensional data clustering for customers with duplicate attribute values [12] –

Customer analysis problem is often involved in high-dimensional data analysis. In addition, the duplicates often disrupt the cluster analysis. This paper focuses on the customer analysis of an educational organization by clustering. Eliminating duplicates is implemented to improve the clustering result.

- Criminal pattern identification based on modified K-means clustering [13] –

Data mining methods like clustering enable police to get a clearer picture of criminal identification and prediction. Clustering algorithms will help to extracts hidden patterns to identify groups and their similarities. In this paper, a modified k-mean algorithm is proposed. The data point has been allocated to its suitable class or cluster more remarkably. The Modified k-mean algorithm reduces the complex nature of the numerical computation, thereby retaining the effectiveness of applying the k-means algorithm.

- Analysis and Clustering of Residential Customers Energy Behavioral Demand Using Smart Meter Data [14] –

Clustering methods are increasingly being applied to residential smart meter data, which provides a number of important opportunities for distribution network operators (DNOs) to manage and plan low-voltage networks. Clustering has a number of potential advantages for DNOs, including the identification of suitable candidates for demand response and the improvement of energy profile modeling. However, due to the high stochasticity and irregularity of household-level demand, detailed analytics are required to define appropriate attributes to cluster. This paper, looks into an in-depth analysis of customer smart meter data to better understand the peak demand and major sources of variability in their behavior.

- Multitask Spectral Clustering by Exploring Intertask Correlation [15] –

Due to the rapid evolution of data on the Web, more emerging challenges have been posed on traditional clustering techniques: 1) correlations among related clustering tasks and/or within individual task are not well captured; 2) the problem of clustering out-of-sample data is seldom considered; and 3) the discriminative property of cluster label matrix is not well explored. In this paper, a novel clustering model, namely multitask spectral clustering (MTSC), to cope with the above challenges is proposed.

V. CONCLUSION

The classic clustering techniques worked with traditional data i.e., simple data like numerical, categorical. Data has evolved from being simple data to object data to complex data. The focus has been captured by Complex Data types like text, stream, temporal, spatial etc., and storage and mining of complex data is challenging. Complex data can also be High Dimensional in nature. The many techniques applied to simple data, do not work on complex data. The classical algorithms have been adapted to handle the new data or specialized techniques have come into existence to work with such data. Current trend is on applying the adapted or specialized techniques on real time data which is typically Complex Data or High Dimensional Data and discover interesting results. These can be further used in building Intelligent Systems which can aid in medical diagnosis, text analysis, genomic data analysis, web analysis, trend analysis, anomaly detection etc.

References

[1] David L. Donoho, "High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality", Aide-Memoire,

Department of Statistics Stanford University August 8, 2000

[2] Michael Steinbach, Levent Ertöz, and Vipin Kumar, "The Challenges of Clustering High Dimensional Data", Chapter - New Directions in Statistical Physics, Pub - Springer Berlin Heidelberg, ISBN 978-3-642-07739-5, pp 273-309

[3] Hans-Peter Kriegel, Peer Kröger, Arthur Zimek, " Clustering High-Dimensional Data: A Survey On Subspace Clustering, Pattern-Based Clustering, And Correlation Clustering", ACM Transactions on Knowledge Discovery from Data (TKDD) Volume 3 Issue 1, March 2009 Article No. 1 ACM New York, NY, USA

[4] Adil Fahad, Najlaa Alshatri, Zahir Tari, "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis", IEEE Transactions on Emerging Topics in Computing (Volume: 2, Issue: 3, Sept. 2014)

[5] H Tao, Y Wang, Z Bu, J Cao, "Discovering Overlapping Communities by Clustering Local Link Structures", Chinese Journal of Electronics, 2017 - IET

[6] G. Shyamala ; N. Pooranam, "A survey on online Stock forum using subspace clustering", IEEE International Conference on Computer Communication and Informatics (ICCCI), 2016

[7] Fahmi Dzikrullah ; Noor Akhmad Setiawan ; Selo Sulistyo, "Implementation of scalable K-Means++ clustering for passengers temporal pattern analysis in public transportation system (BRT Trans Jogja case study) ", IEEE International Annual Engineering Seminar (InAES), Aug 2016

[8] G Wu, H Lin, E Fu, L Wang, "An Improved K-means Algorithm for Document Clustering", IEEE International Conference on Computer Science and Mechanical Automation (CSMA), 2015

[9] S Akter, AS Asa, MP Uddin, MD Hossain, "An extractive text summarization technique for Bengali document (s) using K-means clustering algorithm", IEEE International Conference on Imaging, Vision & Pattern Recognition (icIVPR), 2017

[10] K Mohamed, E Côme, L Oukhellou, "Clustering Smart Card Data for Urban Mobility Analysis", IEEE Transactions on Intelligent Transportation Systems (Volume: 18, Issue: 3, March 2017)

[11] A Nadamoto, S Hanai, H Nanba, "Clustering for Similar Recipes in User-Generated Recipe Sites Based on Main Ingredients and Main Seasoning", IEEE International Conference on Network-Based Information Systems (NBiS), 2016 19th

[12] S Wu, L Fu , "High-dimensional data clustering for customers with duplicate attribute values", IEEE International Conference on Logistics, Informatics and Service Sciences (LISS), 2016

[13] T Aljrees, D Shi, D Windridge, "Criminal pattern identification based on modified K-means clustering", IEEE International Conference on Machine Learning and Cybernetics (ICMLC), 2016

[14] Stephen Haben, Colin Singleton, Peter Grindrod, "Analysis and Clustering of Residential Customers Energy Behavioral Demand Using Smart Meter Data", IEEE Transactions on Smart Grid (Volume: 7, Issue: 1, Jan. 2016)

[15] Yang Yang, Zhigang Ma, Yi Yang, Feiping Nie, Heng Tao Shen, "Multitask Spectral Clustering by Exploring Intertask Correlation", IEEE Transactions on Cybernetics (Volume: 45, Issue: 5, May 2015)