Pedro et al (2014) studied Customer segmentation in a large database of an online customized fashion business with a goal to investigate two different data mining approaches for customer segmentation: clustering and subgroup discovery. In this research the models obtained produced six market segments and 49 rules that allowed a better understanding of customer preferences in a highly customized fashion manufacturer. The outcome of segmentation heavily depended on the input variables, which could be demographic, psychographic, geographic, life-style, etc. Nevertheless, when behavioral data were available concerning what customers purchase, the type of products they prefer, their total expenses, their buying frequency or whether they respond or not to sales promotions, making it possible to implement a more refined segmentation, the current estimation approaches essentially include multivariate analysis. Based on this research framework, the Fashion manufacturers/retailers are closely linked to consumer lifestyle(s). Hence, to enable companies to guide their strategies towards customer satisfaction, models that characterize customer preferences and purchasing behavior are incorporated into their product design processes. Therefore, the most common means of obtaining these models was by using data analysis methodologies. The data used in the research was from Bivolino.com a manufacturer of custom made shirts. Customers can access the Bivolino website or one of its affiliates and purchase a unique shirt by selecting from its many components presented on menus with a large number of alternatives. The total number of customer orders available in this database was 10,775.

Pedro (2014), research presents two algorithms used, which are K-Medoids and CN2-SD. Segmentation criterias that were looked at were Product characteristics, Demographic and biometric (who they are), Geographic (where they live), Psychographic (how they behave) and Behavioral (why they buy). The study was divided into two steps. The first step undertook a clustering analysis based on product characteristics while in the second step, clustering analysis was extended by including customer characteristics in the data. It was discovered that in highly customized industries, such as Bivolino.com, which produces tailor-made shirts, the diversity of products that are sold makes it harder to find clear patterns of customer preferences. The project investigated two different data mining (DM) approaches for customer segmentation: clustering and subgroup discovery. The DM algorithms used, namely K-Medoids and CN2- SD, were valuable instruments enabling a better understanding of consumer tastes and preferences, thus allowing companies to be more efficient and responsive to customer requests and gain a competitive advantage, particularly in highly-customized fashion manufacturing. As hypothesized in the research, these instruments provided different and complementary perspectives on the customers and the products they buy. The results proved useful both for product development and marketing. Despite having presented this case study, it’s believed that the approach described can be useful for marketing in other areas.

A challenge faced by this research was related to the limitations of the method adopted. The K-Medoids clustering algorithm, despite being less sensitive to outliers than K-Means, due to the use of the median instead of the mean, still required the a priori definition of the number of clusters in which, the optimal k number of clusters is known to be a difficult task. Another challenge was presented by the categorical nature of many of the variables used. Nevertheless, the algorithm selected proved to be suitable for the problem under analysis.

From Pedro (2014) application perspective, the study also raised some challenges. It confirmed that in this domain, as in many other fields – such as manufacturing, finance or marketing – the development of DM projects requires a time consuming trial-and error strategy to prepare the data and a fine-tuning of the methods and DM techniques. Moreover, it confirmed that the close involvement of the domain experts is essential for the success of the project.

Stephen L. France and Sanjoy Ghose (2019) presented a review of marketing analytics that primarily covers the topics of visualization, segmentation and grouping, and class prediction. These topics were chosen as not only are they core to marketing strategy and have a long history in academic marketing, but are also of interest to researchers in expert systems, data mining, statistics, and operations research. There is a commonality throughout all three areas. In the 1960s and early 1970s there were a number of papers that took methodology from statistics and psychology and applied it to managerial marketing problems of positioning, segmentation, and response prediction. A core group of researchers in applied psychometrics and measurement, including J. Douglas Carroll, Ronald Frank, Paul Green, and Donald Morrison developed methods in all three of the fundamental areas of visualization, segmentation, and class prediction. The rapid growth of quantitative marketing in this era was spurred on by the availability of computational tools and data, an initiative by the Ford Foundation to equip business faculty with skills in mathematics and quantitative analysis, and the founding of the Marketing Science Institute to support the application of scientific techniques to marketing (Winer and Neslin, 2014)

A paper describing the customer journey and the transformation of customers throughout the life cycle of their use of cashback websites (Maria Teresa Ballester et al, 2018). This paper thus addresses a new and promising research area. By applying the social networks literature to marketing, applying concepts such as loyalty, social networks, and customer evolution and engagement to show that the customer's role depends on the customer's position within the network (Stephen & Toubia, 2009; Zhang et al., 2014). The study also empirically shows that more engaged customers are more transactional, especially in areas where trust is more important (Chen et al., 2014; Pavlou et al., 2007). Finally, they found that in cashback websites, engagement also relates to multi-transactionality, as reported in a more general setting (Bapna and Umyarov, 2015). This analysis has several implications for practitioners, not only in cashback websites, but also in affiliate marketing. The findings show managers how to deal with different customers with different characteristics to strengthen their loyalty and contribution to the brand in a developing area. The findings can yield especially high returns for affiliates in an increasingly competitive environment. The study has some limitations. The findings show a path among clusters that allows practitioners to design marketing strategies depending on the customer's cluster. Practitioners can thus devise strategies for customer acquisition depending on customers' expected value and evaluate customers' online behavior over time.

Yoga Suhas Kuruba and Rasharkin F. Kashef (2021) studied how a customer segmentation model provides decision-makers with an effective allocation of marketing resources and maximization of cross-selling and up-selling opportunities. In the e-commerce industry, customer data are traditionally stored in distributed platforms. It is very costly and might be infeasible to transfer all the distributed data to a single site to apply traditional centralized flat clustering. Thus, distributed clustering provides a global solution to the distributed customer segmentation problem. It is crucial to design a well-structured systematic distributed architecture to perform distributed clustering. Recently, a Peer-to-Peer (P2P) network necessitates the distribution of tasks among peers. However, distributed clustering using the traditional P2P network architectures suffer from communication overhead or inaccurate clustering quality. In this paper, we introduced a novel Multi-tier hierarchical Super-Peer P2P network architecture and algorithm to cluster large-scale distributed customer datasets. Experimental results show that the proposed distributed algorithm outperforms the traditional centralized clustering and proves to be a prudent technique. The proposed distributed approach is evaluated using state-of-art clustering algorithms. The validation metrics include the MSE and the execution time. The proposed MT-SP2P architecture and distributed algorithm achieves a reduction of more than 90% in the clustering error and enhance the clustering speed up to 90% against the centralized approaches. In this paper, we have assessed the performance of the distributed clustering algorithm using various unsupervised machine learning algorithms such as K-means, Fuzzy c-means, Self-organizing Maps, and DBSCN. Our distributed segmentation model provides decision-makers with an efficient provision of the distributed customer segments for better marketing strategies and opportunities. The findings indicated that the proposed model provided better insights and managerial implications with respect to the chosen clustering algorithm and the number of distributed nodes. For large high-dimensional datasets, the computational complexity of deep learning clustering models is very expensive. Thus, future directions for this research include applying deep learning models such as Deep Clustering to evaluate the developed architecture’s performance further. Extended future direction includes investigating the relative impact of the number of nodes initially selected for parallelization, adopting both internal and external quality measures if data labels are provided, and finally evaluating the proposed architecture and distributed algorithm on imbalanced datasets across nodes.

(Avishek Bose et al, 2015) Although hospitality industry is one of the leading business in the world and also increasing its economy steadily, very few research works have been conducted regarding the proper utilization of huge volume of available customer data. This paper provides insights into data clustering features of hospitality big data by analyzing existing density-based clustering algorithms. We have implemented popular density based algorithms, such as DBSCAN, OPTICS, EnDBSCAN, and a few other variants of density-based algorithms, and have provided a comparative performance analysis of these algorithms. Results reveal that EnDBSCAN performs superior than DBSCAN and OPTICS in terms of identifying nested and embedded clusters. Similarly, OPTICS perform better than DBSCAN in identifying adjacent nested cluster for different datasets. However, all of the contemporary clustering algorithms have their limitations in identifying clusters from datasets because of their dependency on input parameters. We can conclude that further research is needed to counter the limitations of existing clustering algorithms. Furthermore, novel clustering algorithms need to be developed for enabling automated recommendation systems for the hospitality industry to improve both customers experience and revenue of the hospitality industry.

In a paper (Deepali Kamthania et al, 2018), an attempt has been made to propose a model for formulating business strategies based on the users' interest and location. The clustering technique has been applied to customer's product-click data for segmentation and PCA technique has been applied to reduce dimensionality. Further, the geographical location has been fetched from an e-commerce website for data visualization. The paper describes the BI tool and a decision process for market segmentation based on user behavior analysis and geographical information. PCA, followed by the application of the k-mode clustering algorithm, has been applied for segmentation. The proposed architecture provides a simplified system for formulating business strategies suitable for the small internet business owners or growing startups. It is a complete toolkit from data cleaning to visualization. The proposed system also identifies the popularity of each product within a time period using interactive visualizations in targeting the unique segments for connecting with potential customers for business expansion.

(Rachid Daoud et al, 2015) presented a case study of applying LRFM (length, recency, frequency and monetary) model and clustering techniques in the sector of electronic commerce with a view to evaluating customers’ values of the Moroccan e-commerce websites and then developing effective marketing strategies. To achieve these objectives, they adopted LRFM model by applying a two-stage clustering method. In the first stage, the self-organizing maps method is used to determine the best number of clusters and the initial centroid. In the second stage, k-means method is applied to segment 730 customers into nine clusters according to their L, R, F and M values. The results show that the cluster 6 is the most important cluster because the average values of L, R, F and M are higher than the overall average value. In addition, this study has considered another variable that describes the mode of payment used by customers to improve and strengthen clusters’ analysis. The clusters’ analysis demonstrates that the payment method is one of the key indicators of a new index which allows to assess the level of customers’ confidence in the company's Website.

(Yulin Deng and Qianying Gao, 2018) the traditional clustering analysis method has obvious lag for the segmentation of e-commerce customers. Therefore, accurate and efficient customer segmentation management should be carried out for the large and complex data information of current e-commerce enterprises, so as to realize customer retention and potential customer mining and promote the efficient development of enterprises. On the basis of customer segmentation theory, for the shortcomings of traditional K-means algorithm, a new SAPK + K-means algorithm based on semi-supervised Affinity Propagation combined with classic K-means algorithm is proposed in combination with AP algorithm, which is applied to e-commerce customers for segmentation management. The results show that when the SAPK + K-means algorithm clusters the iris dataset and the ionosphere dataset, the clustering time is longer than the K-means algorithm and the AP algorithm, but the algorithm error rate in the standard data is significantly reduced and the correct number of clusters can be obtained. The main steps of SAPK + K-means algorithm applied to customer segmentation management including data acquisition, cluster analysis and analysis and evaluation of clustering results. The SAPK + K-means algorithm clusters the data information of an e-commerce customer to obtain four different customer types and proposes corresponding strategies for each type of customer. It is concluded that the SAPK + k-means algorithm can significantly improve the clustering quality of customer data information and improve the effectiveness of activities of e-commerce enterprises.

Yue Li et al (2021) proposed a customer segmentation method based on the improved K-means algorithm and the adaptive particle swarm optimization (PSO) algorithm. The current PSO algorithm can easily fall into a local extremum; thus, adaptive learning PSO (ALPSO) is proposed to improve the optimization accuracy. On the basis of the analysis of population-based optimization, the inertia weight, learning factors, and the position update method are redesigned. To prevent the K-means clustering algorithm from depending on initial cluster centres, the ALPSO algorithm is used to optimize the K-means cluster centres (KM-ALPSO). Aimed at the issue of clustering the actual grape-customer consumption mixed dataset, factor analysis is used to extract numerical variables. We then propose a dissimilarity measurement method to cluster the mixed data. We compare ALPSO with several parameter update methods. We also conduct comparative experiments to compare KM-ALPSO on five UCI datasets. Finally, the improved KM-ALPSO (IKM-ALPSO) clustering algorithm is applied in customer segmentation. All results show that the three proposed methods outperform existing models. The experimental results also demonstrate the effectiveness and practicability of IKM-ALPSO for customer segmentation.

REVIEW OF SOME EXISTING METHODS

**K-Means Clustering**

K-means clustering algorithm is one of the clustering algorithms based on division. It adopts a heuristic iterative process to re-divide data objects and re-update cluster centers. The basic idea of the algorithm is: suppose a set with element objects and the number of clusters to be generated. In the first round, a sample element is randomly selected as the initial cluster center, and the distance between other sample elements and the center point is analyzed the clusters are respectively divided according to the distance. In each of the following rounds, the iterative operation of the above steps is continuously performed, and the average value of the element objects obtained this time is taken as the center point of the next round of clustering until the condition that the clustering center point no longer changes in the iteration process is met.

**Fuzzy C-Means**

Fuzzy logic principles can be used to cluster multidimensional data, assigning each point a membership in each cluster center from 0 to 100 percent. This can be very powerful compared to traditional hard-threshold clustering where every point is assigned a crisp, exact label. This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, summation of membership of each data point should be equal to one. It is an unsupervised clustering algorithm that permits us to build a fuzzy partition from data. The algorithm depends on a parameter m which corresponds to the degree of fuzziness of the solution. Large values of m will blur the classes and all elements tend to belong to all clusters. The solutions of the optimization problem depend on the parameter m. That is, different selections of m will typically lead to different partitions.

**Self-organizing Maps**

A self-organizing map (SOM) is a type of artificial neural network (ANN) that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a map, and is therefore a method to do dimensionality reduction. Self-organizing maps differ from other artificial neural networks as they apply competitive learning as opposed to error-correction learning (such as back propagation with gradient descent), and in the sense that they use a neighborhood function to preserve the topological properties of the input space. SOM was introduced by Finnish professor Teuvo Kohonen in the 1980s is sometimes called a Kohonen map.

**Hierarchical Clustering**

Hierarchical clustering is a method of cluster analysis which builds a hierarchy of data points as they move into a cluster or out of it. Strategies for this algorithm generally fall into two categories:

1) Agglomerative - This clustering algorithm does not require us to prespecify the number of clusters. Bottom-up algorithms treat each data as a singleton cluster at the outset and then successively agglomerates pairs of clusters until all clusters have been merged into a single cluster that contains all data.

2) Divisive - Also known as top-down approach. This algorithm also does not require to prespecify the number of clusters. Top down clustering requires a method for splitting a cluster that contains the whole data and proceeds by splitting clusters recursively until individual data have been splitted into singleton cluster.

Divisive algorithm is also more accurate. Agglomerative clustering makes decisions by considering the local patterns or neighbor points without initially taking into account the global distribution of data. These early decisions cannot be undone. Whereas divisive clustering takes into consideration the global distribution of data when making top-level partitioning decisions.

**Density Based Clustering**

The DBSCAN algorithm is based on this intuitive notion of clusters and noise. DBSCAN is a clustering method that is used in machine learning to separate clusters of high density from clusters of low density. The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points.

There is a relative of DBSCAN, called OPTICS (Ordering Points to Identify Cluster Structure), that invokes a different process. It will create a reachability plot that is then used to extract clusters and although there is still an input, maximum epsilon, it is mostly introduced only if you would like to try and speed up computation time. The other parameters don’t have as big an effect as their counterparts in other clustering algorithms, and are much easier to use defaults.

DBSCAN algorithm requires two parameters:

1) eps : It defines the neighborhood around a data point i.e. if the distance between two points is lower or equal to ‘eps’ then they are considered as neighbors. If the eps value is chosen too small then large part of the data will be considered as outliers. If it is chosen very large then the clusters will merge and majority of the data points will be in the same clusters.

2) MinPts: Minimum number of neighbors (data points) within eps radius. Larger the dataset, the larger value of MinPts must be chosen. As a general rule, the minimum MinPts can be derived from the number of dimensions D in the dataset as, MinPts >= D+1. In this algorithm, we have 3 types of data points. A point is a core point if it has more than MinPts points within eps. A point which has fewer than MinPts within eps but it is in the neighborhood of a core point is border points. A point which is not a core point or border point is noise.

**Affinity Propagation Algorithm**

AP algorithm is a clustering algorithm based on the similarity between N data samples. The AP algorithm doesn’t need to give the initial cluster center or the number of clusters first, but treats all samples as potential cluster centers, called exemplar; it also establishes attractiveness information (that is, the similarity between any two data samples) for each data sample with other data samples with the help of Euclidean distance and stores in similarity matrix which describe similarity between data points. In the AP algorithm, two important parameters are the preference, which controls how many exemplars (or prototypes) are used, and the damping factor which damps the responsibility and availability of messages to avoid numerical oscillations when updating these messages. Different forms of affinity propagation are available like adaptive affinity propagation, partition affinity propagation etc. In order to deal with the clustering time and clustering efficiency of affinity propagation.

COMPARISION SOME CLUSTERING TECHNIQUE

Each clustering algorithm have their own advantages as well as disadvantage with respect to specific situation. K Means is most widely used clustering algorithm for customer segmentation. The K Means require initial number of cluster which is difficult to predict which can affect clustering result. Hierarchical clustering does not require initial number of cluster condition and the time complexity of hierarchal clustering is high and suited for small to medium size dataset. Density based clustering algorithm can be used to find arbitrarily shaped clusters but it not suited for more density difference in data points as well as cluster result efficiency is not good. Affinity Propagation algorithm does not require initial cluster as well as clustering result efficiency is high and clustering time is high which somehow guarantee high clustering efficiency and applicable from small to medium size dataset.

When dealing with large magnitude of data, organizations need to make use of more efficient clustering algorithms for customer segmentation. These clustering model need to possess the capability to process this enormous data effectively. Each of the above discussed clustering algorithms come with their own set of merits and demerits. With different technique pointed above, it came to light that hybrid approach of combining the algorithm can be useful depending upon different situation and the requirement and apply the strategy accordingly. The selection process of clustering technique would require considerable time for studying and implementing as well as processing of data with adequate understanding of goals and apply the algorithm on requirement basis. Hence, it would be helpful for the organization for identifying the distinct group of customers that increases their profit. It also help them in maintaining customer relationship and customer retention by executing different marketing strategies.