PCAEBK-Means Model: A Novel Approach To Determine Number Of Clusters In Auto Component Customers.

C.S Padmasini , Dr K. Shyamala

Abstract: Automotive customer plays an important role in deciding the automotive market and its movement. Based on the Automotive Original Equipment Manufacturer (OEM) requirements, the automotive ancillary manufacturer upgrades their output and satisfies their customer needs. Original Data is categorized into Original Equipment Manufacturing customer supplies, Original Equipment Spare Parts distribution (OES) and to sales and service dealerships. The attributes of the data varies from customer to customer and for a same customer, category to category. The QCD (Quality, Cost and Delivery) criteria's are important to determine customer priorities and group them to find which customer need to be handled on priority basis. PCAEBK-means model is proposed for find the number of clusters which is best suited for this data set. The proposed model comprises the concept of The Principal Component Analysis for dimensionality reduction. Integrated K-means with Elbow method is used to find the optimal number of clusters suitable for the auto component customers.

Keywords: K means Clustering, Elbow method, Principal Component analysis, Eigen values, Auto component, Optimal values

1 INTRODUCTION

Automotive ancillary manufacturers take inputs from Original Equipment Manufacturers (Vehicle Manufacturer), distributors and dealers with respect to vehicle category such as two wheeler, four wheeler, and heavy commercial vehicles. Automotive Ancillary supplier manufacturer supplies products to Original equipment manufacturer, distributors and dealers. Based on their feedback from the customers the products will follow up to their expedition and satisfaction. This deal is considered to be Business to Business category (B to B). The supplier (Automotive Ancillary manufacturer) and the buyer (OEM's) both are into business dealings. They should go in hand in hand to have quality end products. The data set is collected from OEM customers, and its spare parts distributors and dealers. Direct customers play a very limited role so Business to Customers (B to C) is not considered for this study.

2. REVIEW OF LITERATURE

Feng Guo and Huilin Qin [1] propose detailed information about Data mining applied in Customer Relationship Management churn analysis. This paper gave an idea about today's market and how the dimensions about the customer have been changed. The service now a day is all about customer oriented. It explains about how customer is important for business enterprise to survive. Telecommunication data set and applied data mining concepts such as classification and clustering concepts was discussed. Quality of the data is interpreted using Accuracy, Completeness, Consistency Timeliness, Believability, Value added, Interpretability Accessibility characteristics by Burcu Oralhan , Kumru Uyar, Zeki Oralhan [2] Gaurav Gupta and Himanshu Aggarwal [3] explain about operational management needs analytical management with predictive data mining models. The route to a successful business requires a marketing manager who understands his customers and their requirements and implements data mining with a good unique model. This paper has explained about clustering methods and unsupervised learning. PCAEBK-means model proposes to identify the suitable optimal number of clusters.

The data set has 24 attributes which are broadly classified into quality, cost, delivery and new product development. This data set consists of categorical and numeric data. Though there are many data types available, more than 80% of data are Categorical data in this automotive data set. Principal component analysis (PCA) is mathematical procedure that maps a number of correlated variables into a smaller set of uncorrelated variables called as Principal component analysis[16]. K means clustering algorithm is very popular and traditional clustering method integrated with Elbow method to find best cluster formation. The result of graph generated in Elbow method shows suitable and best k value for the automotive customer data set. Clustering algorithms are estimated based on accuracy metrics.

Clustering is grouping data objects based on various categories. It explains about Partitioned clustering, Hierarchical clustering, Density based clustering by Saroj , Tripti Chaudhary [4].Apurva Juyal, Dr. O. P. Gupta [5] explains about hierarchical and partitioned clustering methods in detail. Partition clustering is based on center point of the clusters. Explanation of pattern extraction for customer data is done which is important in business process. Objective of the paper is to identify high profit, high value and low risk customers by clustering techniques. Here demographic clustering is used by Dr. Sankar Rajagopal[6]. This Author proposes a comparative analysis between PCA with Kmeans and Fuzzy Cmeans clustering for customer segmentation, ahmida Afrin, Md. Al-Amin, Mehnaz Tabassum [16]. With the research papers listed above, Hybrid PCAEBK-means proposed model can be formulated for customer segmentation. To value the quality of the clustering various techniques such as accuracy can be implemented. To find the optimal value of number of clusters Elbow method may be used.

3. PROPOSED WORK

The proposed novel model Hybrid PCAEBK-means gives a different perspective to handle customers to find better business solutions. This novel framework has various layers which will be explained in detail and to know the importance of each and every layer. The first step towards customer data set is preprocessing the data into a meaningful form. The data set may have in complete data and missing values. This is handled in a professional way. A user defined method is implemented for converting categorical values to numerical values as K means support numerical data. The Second step is to reduce the dimensionality by PCA algorithm. The next step will be Implementation of Integrated K means with Elbow method for unsupervised learning.

3.1 Preprocessing

The Automotive customer data is preprocessed to obtain final data with more clarity. If the customers have given the values for all the attributes, then the data will be complete and can be considered whereas certain data may be incomplete and has below 50% of data then those can be avoided and cannot be considered for analysis. If few data are only missing, the customer data will be considered by applying the mean value of the existing data. Since the data set has more than 20 attributes for customer requirement and satisfaction criteria, the analysis is very difficult to validate the data. Most of the attributes are categorical data type in this data set. These data should be converted into readable format which taken to next level of the work.

3.2 Principal component analysis (PCA)

Principal component analysis (PCA) is probably the best known and most widely used dimension-reducing technique [7]. As an unsupervised learning method, PCA has numerous applications in various fields. The success of PCA is due to the following two important optimal properties: 1. Principal Components sequentially capture the maximum variability among X, thus guaranteeing minimal information loss; 2. Principal Components are uncorrelated [8].It is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. This transformation is defined in such a way that the first principal component has the largest possible variance and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set [17].

Algorithm: Principal component Analysis
Input: Auto Ancillary customer data set
Output: Dimensionality reduction

1. Compute the mean of every dimension of the whole dataset
2. Compute co variance matrix for the data set.
3. Calculate Eigen vectors and corresponding Eigen values
4. Sort the Eigen vectors and choose the largest Eigen values to form a matrix
5. Transform the samples into a new space

The equation (1) calculates the co variance matrix of the data set. it is the intermediate step of principal component analysis followed by finding Eigen values and Eigen vectors.

$$\mathrm{cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(Xi - x')(Yi - y') \quad \text{----}$$

3.3 Determine the optimal K value using Elbow method

To find the optimal number of clusters there are mainly three standard methods used. 1. Silhouette method 2. Elbow method 3. Gap statistic method. Out of three to find the optimal number of clusters, elbow method is used.
Elbow method is a method which looks at the percentage of variance explained as a function of the number of clusters. This method exists upon the idea that one should choose a number of clusters so that adding another cluster doesn't give much better modeling of the data [14]. It is a visual method. The idea is that Start with K=2, and keep increasing it in each step by 1, calculating your clusters and the cost that comes with the training [15].This method exists upon the idea that one should choose a number of clusters so that adding another cluster doesn't give much better modeling of the data [16].

Algorithm: Elbow Method to determine optimal K value for the data set
Input: Auto Ancillary customer data set
Output: Optimal K value is found

1. Initialize k=2
2. Start the method
3. Increment the value of k by 1
4. Find the suitable K value for the data set
5. End

3.4 Data clustering

The K-Means clustering algorithm is proposed by Mac Queen in 1967 which is a partition-based cluster analysis method. It is used widely in cluster analysis for that the K-means algorithm has higher efficiency and scalability and converges fast when dealing with large data sets [11]. K means clustering is one of the famous and mostly used standard clustering algorithms for categorical data. K means clustering is a non hierarchical clustering procedure in which items are moved among sets until the desired set is reached [9]. The partitioning of data set is such that the sum of intra cluster distances is reduced to an optimum value [10]. K means clustering is used to cluster the data with possible meaningful k values such as k as 2, 3 4, 5. If K value is 1 the customer needs cannot be justified. If K value is above 5, segmenting the customer and finding the solutions is very difficult due to multiple attributes complexity. The concluded K values can be either 2 or 3 or 4 or 5. The results are generated to find out the best clustering for the data set.

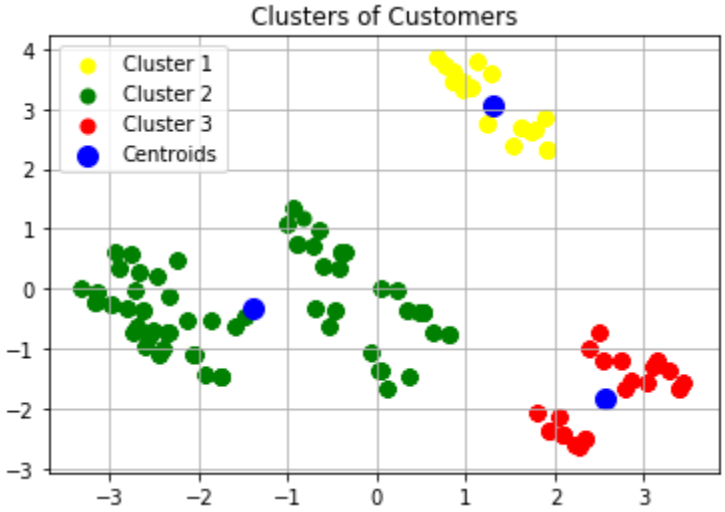

***Figure 2**: Clustering the Automobile customer data*

This novel method based on Figure 2 shows that clustering optimal value can be 3 and well suited for auto ancillary products

4. RESULTS AND DISCUSSIONS

This algorithm is proposed to find the optimal k clusters suitable for the data set. Preprocessing is done to make the data complete. if the data given by the customer is in complete the customer information is not considered as it produces incorrect data. If the customer has left only few criteria the data is considered and find the mean of the values and give the new value for the customer data. This after preprocessing the data set will be complete and ready for next stage in the algorithm. PCA algorithm is used for Dimensionality reduction. This is very important part of this work because the data set has more than 20 attributes, to relate each and every attribute and to correlate dimensionality reduction, PCA techniques are used. K means and Expedition maximization algorithms are used for clustering the data set. Optimal K value is found with the help of Elbow method. The graph depicts a direction change in the line. The bend in the graph says the suitable and optimal K value for the data set. As a result K=3 is suited for this data set.

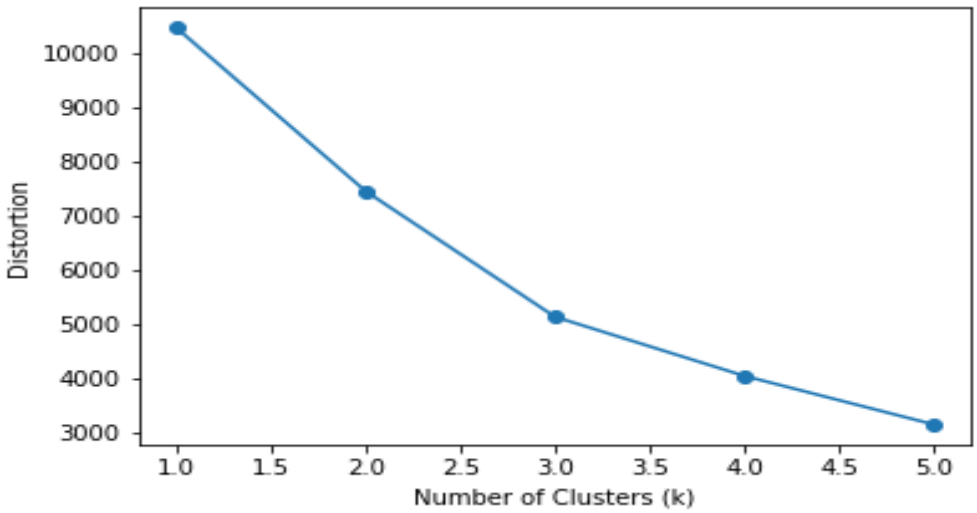

***Figure 3:** To find K value using Elbow method*

The result of Elbow method is shown in Figure 3 and output is generated. The output shows a bend in the value of 3. This depicts the suitable K value for this auto ancillary product dataset, the data set will be clustered as 3 groups as Excellent, satisfied, not satisfied categories.

5. CONCLUSION

Automotive ancillary manufacturers play a vital role in dealing with customers who have to provide their ultimate products or units or vehicles to their respective customers, OE Spare Distributors and dealers, the best of products for the value of money they spent on it. The goal of making PCAEBK-Means is to find optimal K value for the data set. Graph of the Elbow method depicts the automotive data set should be clustered as 3 groups. These three groups define whether the customer feels good, better and bad about the auto ancillary products.

REFERENCES

[1] Feng Guo and Huilin Qin, Data Mining Techniques for Customer Relationship Management, IOP Conf. Series: Journal of Physics: Conf. Series 910 (2017)

[2] Burcu Oralhan , Kumru Uyar, Zeki Oralhan ," Customer Satisfaction Using Data Mining Approach, International Journal of Intelligent Systems and Applications in Engineering

[3] Gaurav Gupta and Himanshu Aggarwal, "Improving Customer Relationship ManagementUsing Data Mining", International Journal of Machine Learning and Computing, Vol. 2, No. 6, December 2012

[4] Saroj , Tripti Chaudhary, "Study on Various Clustering Techniques "International Journal of Computer Science and Information Technologies,Vol. 6 (3) , 2015, 3031-3033

[5] Apurva Juyal, Dr. O. P. Gupta , " A Review on Clustering Techniques in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 7, July 2014

[6] Dr. Sankar Rajagopal," Customer Data clustering using data mining techniques", International Journal of Database Management Systems (IJDMS) Vol.3, No.4, November 2011.

[7] Ian Jolliffe, International Encyclopedia of Statistical Science" 2011, springerlink

[8] https://www.cc.gatech.edu/home/isbell/classes/reading/papers/sparsepc.pdf

[9] J. Han, M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, pp. 401-404, 2007.

[10] D. Steinley, M.J. Brusco, "Initializing K-Means Batch Clustering: A Critical Evaluation of Several Techniques", J. Classification, vol. 24, pp. 99-121, 2007.

[11] Juntao Wang ; Xiaolong Su, " An improved K-Means clustering algorithm", IEEE 3rd International Conference on Communication Software and Networks 2011.https://ieeexplore.ieee.org/abstract/document/6014384

[12] https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm

[13]https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm#Applications

[14] Purnima Bholowalia,Arvind Kumar, " EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN",International Journal of Computer Applications,Volume 105 – No. 9, November 2014

[15] Trupti M. Kodinariya , Dr. Prashant R. Makwana, "Review on determining number of Cluster in K-Means Clustering", International Journal of Advance Research in Computer Science and Management Studies, Volume 1, Issue 6, November 2013.

[16] Fahmida Afrin, Md. Al-Amin, Mehnaz Tabassum, "Comparative Performance Of Using PCA With KMeans And Fuzzy C Means Clustering For Customer Segmentation" International Journal Of Scientific & Technology Research Volume 4, Issue 10, October 2015 ,ISSN 2277-8616

[17] https://en.wikipedia.org/wiki/Principal_component_analysis