Prediction of Customer Churn in Telecom Sector using Clustering Technique

Vallabhaneni Renuka Devi, G. Bharathi, G.V.S.N.R.V. Prasad

Abstract: ***These days the data is producing at an incredible rate. Handling and analyzing such a big data in a specific time is the main challenge today. Clustering is majorly familiar with analyzing the data visually and used for efficient decision making process. Clustering is broadly used in a range of applications like education, field of computer science, marketing, insurance, surveillance detection, fraud detection and scientific discovery to mine the functional information from the data. This paper concentrates on the unsupervised learning k-means clustering algorithm to perform the analysis on churn prediction on telecom sector. The selection of distance measures and the category of data that a clustering algorithm cans effort is a decisive step in clustering. It defines how two elements are resemblance with each other and how this resemblance will impact the outline of the clusters. Another foremost difficulty in clustering process is to determine the goodness or validity of the cluster. Hence this paper discusses and addresses the different issues with K-means clustering. Experimentation was done on china telecom data to identify analogous group of clients who more likely to prone from the services is a major task. The results were analyzed to identify best feature, distance measures and validity indices to get qualitative clusters.***

Keywords: clustering, unsupervised learning, K-Means, Validity indexes.

I. INTRODUCTION

Clustering is the procedure of extracting set of items or objects from the given data points rely on their features and then aggregating them according to their similarity. In data mining, clustering [1] methodology first step is to partition the given data and then implements a precise join algorithm that is suitable for the data analysis in universal space.Cluster of information objects from the given data points are treated as a group at an instance. While performing Clustering, first split the arrangement of data objects into groups dependent on information closeness and assigns out the data points to the closest groups. Well known structure of Cluster comprises of groups with little separation between the group of individuals, thick regions of the universal data space and in terms of measurable conveyances.

Clustering accordingly can be figured as multi-target optimist problem. The proper clustering technique and parameter or bound evaluation rely on the specific domain data set and deliberate utilization of the conclusions. Various applications of clustering.

A. Medicine Field

The specialist would identify side effects, such as psychology, tension, anxiety, depression etc. The cluster examination can recognize groups of patients that have similar side effects.

B. Business and Marketing Field

What are customer segments? To response this inquiry an economic specialist may direct an overview for requirements, attitudes, socioeconomics, demographics, and actions of customers. The researcher at that point then use cluster examination to identify analogous group of clients that have same needs, attitudes and mentalities. Market research: Cluster investigation can be performed on statistical surveying issue. Economic scientists perform the statistical cluster analysis to isolate the overall public of buyers into market portions. Market research helps to enhance the relationship between diverse groups of consumers/potential customers.

C. Education Field

What are the student groups that require particular consideration? Specialists may compute psychological, inclination, aptitude and accomplishment characteristics among the students. Cluster investigations can identifies the homogeneous groups that exist between students (for instance, the achievers in all students or the subjects that surpass in certain subjects but fail in others).

D. Biology Field

In this field cluster investigation is mainly used in categorization of species. Analysts can gather a data set of divergent plants and can note diverse characteristics of their phenotypes. A cluster examination can group those observational records into a progression of clusters and assemble scientific classification of gatherings and subgroups of practically equivalent to plants.

Revised Manuscript Received on August 20, 2019.

* Correspondence Author

Vallabhaneni Renuka Devi*, M.Tech Student, CSE, Gudlavalleru Engineering College, Gudlavalleru,(A.P), India, (Email:- renukavallabhaneni@gmail.com)

Mrs. G. Bharathi, Professor, CSE, Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh (A.P), India, (Email:- gbharthi@gmail.com)

Dr. G.V.S.N.R.V. Prasad, Professor and Vice principal of academics CSE, Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh (A.P), India, (Email:- gutta.prasad1@gmail.com)

E. World Wide Web

In social network analysis, clustering derives the acquainted results with communities inside extensive groups of individuals. Search result grouping: While performing intelligent grouping of files or documents and websites, clustering can generate a more pertinent and accurate set of search results.

F. Computer Science Field

Software evolution: Clustering plays an imperative job in evolution of software by helping to inheritance legacy properties in code. Image segmentation: Clustering isolate a digital image into particular locales for outskirt or protest acknowledgment.

G. Social Science Field

Crime analysis is the main areas of social science. The areas that are more noteworthy occurrences of specific incidents of wrongdoing are identified by cluster analysis[7]. By knowing about these "hot spots", where a same crime has been happening for many times over a period of time, this results in utilization of law enforcement resources more successfully.

H. Others

Clustering finds many applications inrobotics to identify the anomalies in sensor information, Climatology to discover barometrical patterns and in Petroleum geology.

From all of these applications focus on churn prediction in telecom sector[11] which comes under the business and marketing field to cluster the customers.

II. VARIOUS CLUSTERING ALGORITHMS

The primarily important clustering techniques that can be performed in big data[3] are Partition based clustering algorithms[2], Hierarchical, Grid based, Model based and Density based algorithms. The option of clustering algorithms [8] always relies on the characteristics of the given data and the desire to accomplish task with that data set.

A. Partitioning Based Clustering

Partitioning clustering contains clustering methods that are used to classify the observations or records within a data sets into multiple clusters based on their resemblance. Clustering requires a predetermined no of clusters to specify a priority. Squared error function is used as an objective functions as a measure in the data partitioning optimization process. Partitioning clustering executes the iterative process to optimally declare the cluster centers and number of clusters for the given clustering algorithm. These algorithms require the data scientist to denote the number of clusters that is k value to be generated. For example by constructing a k partitions (k <= n) and then there is a need to evaluate them by using some decisive factor based on the algorithm, e.g., minimize the sum of squared errors which includes the cluster withinness by following some rules such as each cluster should hold at least one object and at most each object should belong to one group only. Some of the important partitioning algorithms used are K-means, FCM and K-Medoids etc.

B. Hierarchical Clustering

Hierarchical clustering [3]does no longer state the number of clusters, and the output is unbiased of the initial circumstance. In hierarchical clustering approach, initially start with one item and then successively or iteratively merges the neighbour items based on some type of distance criteria(should be minimum). Dendograms are familiar to represent the clusters. Some of algorithms are BIRCH, Chameleon, CURE, SNN etc.

C. Density Based Clustering

Density based clustering algorithms clusters the information items based totally on their regions of boundary, density and connectivity. Density-based clustering is superior for identifying clusters of uninformed shapes against the outliers. GDBSCAN, DBSCAN, OPTICS and DENCLU are some of density based clustering techniques.

D. Grid Based Clustering

Grid based clustering [3]divides the records into number of grids. It is finished by specifying the information set as soon as to figure out the statistical values for the grids with a fast dispensation time.

III. K-MEANS

K-means Clustering Method: Fromthe various clustering algorithms let us select the K-Means algorithm to perform the analysis for a specific application. K-Means algorithm firstly start with partition of given set of items or objects into k number of non-empty subsets. Identify the k cluster centroids of the present partition. Assign each point in a specified data set to a particular cluster. Calculate the distances from every point and assign data point to the closest or nearest cluster, pick the minimal distance from centroid. After re-dispensing the points, discover the center of the new cluster formed. Reduce the SSE with post processing: The objective is to diminish the SSE and one apparent way to do this is to increase the variety of clusters. Two techniques that reduce the full SSE via growing the quantity of clusters are to split the cluster or to introduce a new cluster centriod. Applications of K-Means **:** Pricing, loyalty and behavior Segmentation, Branch Geo Segmentation, Category Segmentation, Customer Need Segmentation, Donor Segmentation, Server Clustering, Healthcare Fraud Detection Segmentation.

IV. APPLICATION OF K-MEANS CLUSTERING ON TELECOM DATA:

Churn prediction[6] is vital for business as it helps you to spot customers who are probable to stop the subscription of service or product. Churn prediction[5] is very helpful for customer preservation and can predict in advance the customers those are at risk of separation. Telephone service companies [7] frequently use

customer abrasion analysis and the customer erosion rates as one of their key commerce metrics because the price of retain an alive customer is outlying fewer than acquire a new one. Companies from this sector frequently contain client examination branches which effort to succeed back falling clients, because improved long-term customers can be significant much supplementary to a company than newly recruit clients. Consider the china telecom churn [5] data set with attributes state, account length, area code, international plan, total international calls total day minutes, total day calls, total day charge, total eve minutes, total night charge, total evening calls, total evening charge, voice mail plan, total night minutes, total international charge, total night calls, total night charge, total international minutes, total international calls, customer service calls, and churn with 3333 records. **Target Variable is considered as c**hurn: if the customer has churn (1=no; 2 = yes). Perform the cluster analysis to predict the areas that are previously mentioned or the features of customers that are more prone to churn. So it is necessitate designing a K-Means clustering algorithm to perform the churn prediction analysis. While performing K-means, there are various issues such as size of the data that can be choose to perform cluster analysis, features that are to be selected from the considered input data (dimensionality reduction), selection of number of clusters i.e. k value and the convergence criteria for K-Means algorithm. Let us discuss these issues for k-means technique in following sections. Following sections also shows the experimental analysis for churn prediction performed by using K-Means algorithm. By this analysis part, the best solution or the cause for the churn detection [5]in telecom data sector.

V. DATA PRE-PROCESSING AND ISSUES FOR K-MEANS CLUSTERING:

Data preparation for k-means clustering includes following steps:

A. Normalization

It is a pre-processing technique which can apply to a data set before performing cluster analysis. Generally normalization is used for scaling the values in the range of [0, 1] for all the values of columns for specific data. By applying normalization technique, one can significantly see the accuracy in the results. Particularly by using normalization technique clustering results can be obtained with minimizing SSE.

a. Min-Max Normalization: is primarily used normalized technique to normalize the feature of data. Formula for min-max normalization is: Y represents the column of data set.

$$Z = \frac{Y - MIN(Y)}{MAX(Y) - MIN(Y)} \quad (1)$$

b. Z-score Normalization: It describes the deviation of difference between particular item to a standard deviation and mean of the observation of a column. Where SD represents standard deviation.

$$Z = \frac{Y - MEAN(Y)}{SD(Y)} \quad (2)$$

B. Feature Selection for K-Means Clustering

If the data set contains more attributes [4] it is difficult to scale. So feature selection will reduce the time complexity.It is worthwhile to weight variables, unavoidably a few factors will contain more data about the clusters than others. Feature selection**:** there are two mainly two types of variable or feature assortment methods for clustering they are: filter methods and wrapper methods. Filter methods mainly concentrate on single factor analysis. The predictive capability of every entity variable is evaluated. It contains some methods like: Information gain, correlation measure, Gain ratio and symmetrical relevance. In wrapper methods the predictive and analysis capability of the variable or feature is evaluated jointly with comparison of other variables from data set. It contains some methods like subset selection, backward subset elimination and another important one is forward subset selection methods.

a. Information Gain: generally entropy characterize the impurity measure of a randomly collection of examples. Information gain is the ordinary lessening in entropy causes by dividing the records according to given attribute. E represents entropy and IG represents information gain. Select the attributes with highest IG than a threshold value.

$$ENTROPY(A) = \sum_{I=1}^{K} -C_I \log_2 C_I \quad (3)$$

$$IG = E(class) + E(attr) - E(class, attr) \quad (4)$$

b. Gain Ratio: Select the attributes with maximum Gain ratio than a threshold value.

$$\text{gain ratio} = \frac{E(class) + E(attr) - E(class,attr)}{E(attr)} \quad (5)$$

c. Symmetric Relevance: Select the attributes with highest symmetric relevance than a threshold value.

$$SR = 2 * \frac{E(class) + E(attr) - E(class,attr)}{E(attr) + E(class)} \quad (6)$$

d. Subset Selection [10]: Start with fitting all combination of models with the given variables or features that have k predictors, out of these k predictors select the p predictors. This selection is done by selecting best model from model (1), model (2)... model (p). Use the RSS, Adjusted R Square to fit the model. RSS should be min for selected model.

e. Forward Selection: Forward subset selection [10] is continuously performing iterations have no feature or variable in the model initially. In each process of iteration, adding up the feature which most excellent improves the representation till an adding up of a new variable do not progress the best performance of the constructed model. Select the attributes with min forward selection than a threshold value.

f. Backward Elimination: In backward subset elimination [10], will commence with all the variables or features and then remove the slightest important feature from iteration, which improve the performance of the considered model. Select

the attributes with min backward elimination than a threshold value.

g. Correlation: It describes how well one attribute or feature is strongly related with class label and the relation between the other variables must be low. Formula for correlation analysis is:

$$COR(Y,X) = 1 - \left(\frac{\left(\sum_{I-1}^{N}(Y_I - \bar{Y}_I)^2(X_I - \bar{X}_I)^2\right)}{\sqrt{\sum_{I-1}^{N}(Y_I - \bar{Y}_I)^2 \sum_{I=1}^{N}(X_I - \bar{X}_I)^2}}\right) \quad (7)$$

The y and x from above formula denotes the two columns in the considered data set. Select the attributes with maximum Correlation than a threshold value.

C. Distance Metrics:

a. Euclidean distance: $D(Y,X) = \sqrt{\sum_{I=1}^{N}(Y_I - X_I)^2}$

b. Manhattan distance: $D(Y,X) = \sum_{I=1}^{N}|(Y_I - X_I)|$

c. Pearson correlation distance:

$$D(Y,X) = 1 - \left(\frac{\left(\sum_{I-1}^{N}(Y_I - \bar{Y}_I)^2(X_I - \bar{X}_I)^2\right)}{\sqrt{\sum_{I-1}^{N}(Y_I - \bar{Y}_I)^2 \sum_{I=1}^{N}(X_I - \bar{X}_I)^2}}\right) \quad (8)$$

d. Eisen cosine correlation distance

$$D(Y,X) = 1 - \left(\frac{\sum_{i-1}^{N} Y_I X_I}{\sqrt{\sum_{I-1}^{N} Y_I^2 * X_I^2}}\right), \quad (9)$$

e. Kendall correlation distance:

$$D(Y,X) = 1 - \left(\frac{N_D - N_C}{\frac{1}{2}N(N-1)}\right), \quad (10)$$

Thetypes of distance measureswhich are chosen are depend on data set. Among these distance metrics, Euclidean measure is considered for applying K-Means algorithm.

D. Selection of k Value: It is thedecisive step in theclustering process to select the best k value. Some methods are:

a. Elbow Method: This method concentrates on reducing the withinness of cluster variation. It computes the k-means algorithm for different values of k. For every k, computes the WSS.

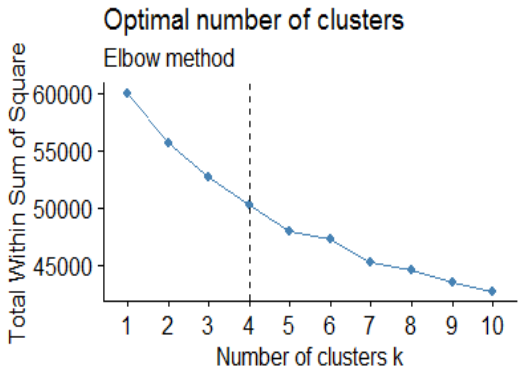

Fig.1. K value selection using elbow method

Then design the plot of WSS (Within sum of squares)acc to the figure of groups of clusters. From the figure 1 elbow method suggests that the optimal clusters are four.

b. Silhouette Method: It measures the excellence of clustering. That means it determines how finely each item classifies within its desired cluster.

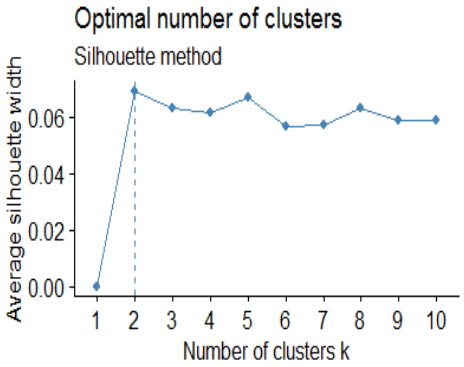

Fig.2. K value selection using silhouette method

From the figure 2 Silhouette methods suggests that the optimal clusters are two.

c. Gap Statistic Method: It compares the total withinness cluster differentiation for various values of k. For each variable (x_i) in data set computes its range [min (x_i), max (x_j)] and generate the n values uniformly from the interval min to max.E_n is determines via bootstrapping. W_k is the clustering point that is farthest from uniform distribution of points.

$$Gap_n(k) = E_n * \log(w_k) - \log(w_k) \quad (11)$$

From the figure 3Gap statistic method suggests that the optimal clusters are two.

Elbow method chooses the k=4 (fig 1) which optimizes the result, Silhouette method (fig 2) and Gap statistic method (fig 3) chooses the k=2 which optimizes the result. The results of above three methods show that two number of clusters are more appropriate and qualitative.

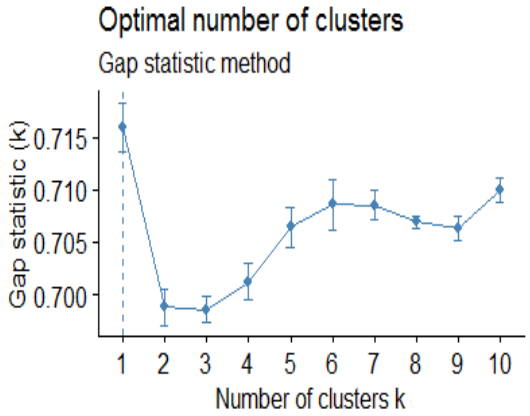

Fig.3. K value selection using Gap statistic method

VI. VALIDITY OF CLUSTERING

Some of different elements of cluster validation are figuring out the clustering tendency of a set of records, comparing the set of result of a cluster assessment to externally recognized consequences, e.g., to specify the class label for the given records, comparing the outcomes of a cluster examination vigorous the statistics without orientation to outside records.

A. Internal Cluster Validation: inner cluster validation[12] uses the internal facts of the clustering procedure to estimate the goodness or validity of a clustering group without reference to external statistics [10]. It can be extensively utilized to approximation k value for given facts contain and the best clustering algorithm without any outside information.

a. Connectivity[9]: measures how items are placed within the equal cluster compared with their nearest neighbours inside the given data area. The connectivity degree continually lies between 0 and infinity and to this price ought to be minimized.

b. Silhouette Index[9]: For any of the given cluster, (j = 1,,), the silhouette system assign Xj to the ith sample of a first-rate measure, (i = 1, , … m), referred to as the silhouette thickness(width).

$$SI(d) = \frac{y(d) - x(d)}{MAX\{x(d), y(d)\}} \qquad (12)$$

Where (d) is the common distance among the ith sample of data.

c. Dunn Index[9]**:** Dunn is used for measuring the quality of the cluster:

$$dunn = \min\{\frac{d(c_j, c_i)}{\max(d(y_k))}\} \qquad (13)$$

Where (ci, cj) define the inter cluster distance among the given clusters Y_i and X_j; (Y_k) represents the intra or inner cluster distance of cluster (Y_k) and c is the variety of clusters for the specified dataset. If the values of index Dunn [10] are huge then, it corresponds to appropriate clustering solution.

B. External Cluster Validation: external cluster validation[12] performs evaluating the results of a cluster evaluation to an externally recognized end result, such as externally provided magnificence labels. Then it measures the degree to which cluster labels match among the externally detailed elegance labels.

a. F-measure[10]**:** is outside validity indexing degree, which mixes the precision and take into account concepts from facts retrieval. Recall and precision of that cluster for every record is calculated as:

$$Recall(j,i) = \frac{n_{ij}}{n_j} Precisioon(j,i) = \frac{n_{ij}}{n_i} \qquad (14)$$

Were n_{ij} is the wide variety of objects of the class i that belong to cluster j, $F - Measure$ of cluster's j and class i is given by means of the following equation:

$$F(j,i) = \frac{2Recall(j,i)Precision(j,i)}{Precision(j,i) + Recall(j,i)} \qquad (15)$$

Table - I: Matrix

MATRIX	*Same cluster*	*Distinct clusters*
Same class	true positive(TP)	false negative(FN)
Diff classes	false positives(FP)	true negative(TN)

From table I TP denotes true positive, it occurs while taking a decision to assign two similar points to the same cluster, TN denotes a true negative; it occurs while taking a decision to assign two unequal or dissimilar points to the distinct clusters. FP denotes false positive, it occurs while taking a decision to assign two unequal or dissimilar points to the identical cluster. FN denotes the false negative it occurs while taking a decision to assigns two alike or similar points to distinct clusters.

b. $\text{Rand Index} = \left(\frac{TP+TN}{TP+FP+FN+TN}\right)$ (16)

For good cluster this index should be higher.

c. $ADI = \left(TP - \frac{PRODCOMB}{MEANCOMB - PRODCOMB}\right)$ (17)

Adjusted random for good cluster this index should be higher. Where product represents (tp +fp)*(tp +fn) and mean represents (prod)/2

d. $JACCARD\ INDEX = \left(\frac{TP}{TP+FP+FN}\right)$ (18)

For good cluster this index should be higher.

e. $FMI = SQRT[\left(\frac{TP}{(TP+FP)}\right) * \left(\frac{TP}{TP+FN}\right)]$ (19)

Fowlkes mallows index [10]for good cluster this index should be higher.

f. $MIRKINNMETRIC = (2 * (FP + FN))$. (20)

For good cluster this index should be lower.

g. $\mathbf{PURITY} = \left(\frac{TP}{TP+TN+FP+FN}\right)$ (21)

For good cluster this index should be higher. Purity measure is very a whole lot just like entropy. For every one cluster, the purity is the range of objects in j with recognize to the elegance label i. The normal cleanliness of the given clustering solution is received as a subjective amount of the entity cluster purities within the given information.

$$\mathbf{ENTROPY}(A) = \sum_{I=1}^{K} -C_I \log_2 C_I \qquad (22)$$

For good cluster this index should be higher.

h. $\mathbf{NMI} = \left(2 * \frac{MI(X,Y)}{[E(X)+E(Y)]}\right)$ (23)

Normalized mutual information for good cluster this index should be higher. By performing external validation indexes on the churn data set will return the k value that best suitable for the problem considered. These indexes also help to determine the good qualitative clusters those can be constructed from the given data.

VII. EXPERIMENTAL RESULTS

Table – II: Results of normalization

		K=2	K=3	K=4	K=5
Without normali-zation	WSS	368227	3244971	2948778	2718286
	BSS	7.3176	7.3176	7.31762	7.31762
Min-max normaliz ation	WSS	65580.	63480	62380	66879
	BSS	7.3176	7.3176	7.31762	7.31762
Z-score normaliz ation	WSS	55580.	52569	49932.8	47865.8
	BSS	7.3176	7.3176	7.31762	7.31762

Table II shows the WSS (within Sum of squares) and BSS (between sums of squares) of the cluster formed by selecting various values of k before normalization and after performing normalization. Above table shows the results for various clusters by applying normalization and without applying normalization. It also shows that withinness is decreased enormously and betweenness increased after applying normalization. Out of these two normalizations by applying Z-Score normalization technique well generates the separated and well dense clusters.

A. Feature Selection:

Table – III: Results of feature selection

Feature selection	**Threshold value**		**K=2**	**K=3**	**K=4**
Without feature selection	**18 features selected**	WSS	55580	52572	49897
		BSS	4965	7403	10078
Information gain	**0.00467 (7 features selected)**	WSS	9521	5359	3364
		BSS	4281	7244	9864
Gain ratio	**0.00467 (10 features selected)**	WSS	29008	26042	23388
		BSS	4311	7279	9922
Symmetric relevance	**0.00467 (10 features selected)**	WSS	29008	26042	23388
		BSS	4311	7279	9922
subset selection	**0.05821 (7 features selected)**	WSS	19972	16966	15476
		BSS	3351	6357	7847
Forward selection	**0.05821 (6 features selected)**	WSS	16641	13635	12145
		BSS	3350	6356	7828
Backward elimination	**0.05821 (6 features selected)**	WSS	16641	13635	12163
		BSS	3350	6356	7828
Correlation	**0.05821 (9 features selected)**	WSS	17656	14658	13568
		BSS	4320	5962	6521

Table III shows the comparison of results for various clusters with and without applying feature selection techniques. If feature selection is not applied all the eighteen attributes from the given dataset will be selected. Then K-means constructs a cluster by taking all these attributes as an input which takes more time to construct a cluster. Moreover some attributes may be irrelevant to construct a cluster. So adapting feature selection methods gives better results with less time complexity. Feature selection techniques are classified in to two types such as filter methods and wrapper methods as mentioned earlier. For filter methods consider 0.00467 as a threshold value. Then information gain selects the best 7 features that can derive more quality cluster. Gain ratio and Symmetric relevance selects 10 features. For wrapper methods 0.05821 is selected as a threshold value. Then subset selection measure selects 7 features, Forward selection and backward elimination measures selects 6 features and correlation selects 9 features. From above information gain gives the smallest withinness and largest betweenness at k=5 for the selected threshold value. This represents by selecting the information gain as a feature selection for the churn data set, it retrieves the dense and well separated cluster with less time complexity.

B. Clustering Result:

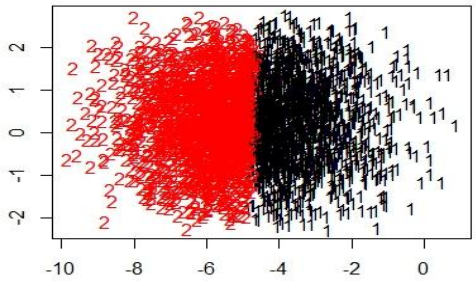

Fig.4: Results of clustering

This figure IV shows two clusters. Cluster 1designates the customers those are not churn and cluster 2 designates the customers those are prone to churn.

C. Internal Cluster Validation

Table –IV: Results of internal cluster validation

	k=2	K=3	K=4	K=5	Best
Connectivity	32.74	48.478	70.928	84.594	K=2
Dunn index	0.0230	0.0290	0.0211	0.0283	K=3
Silhouette	0.4728	0.3606	0.3376	0.3411	K=4

For determining the best qualitative cluster it is required to choose a good k value. Table 4 shows if Connectivity is chosen as a internal validity index, it forms the qualitative cluster at k=2(at which the connectivity index is less for k value). If Dunn index is chosen as an internal validity index, it forms the qualitative cluster at k=3(at which the Dunn index is more for k value). If Silhouette index is chosen as an internal validity index, it forms the qualitative cluster at k=4(at which the Silhouette index is less for k value).

D. External Cluster Validation:

Table – V: Results of external validation indexes

	K=2	K=3	K=4	K=5	Best
Rand index	0.5022	0.416504	0.410138	0.369445	K=2
Adjusted rand index	0.0045	0.001049	0.033417	0.021584	K=4
Jaccard index	0.4310	0.300690	0.473248	0.209893	K=4

Fowlkes mallows	0.6150	0.501197	0.482148	0.418048	K=2
Merkinn metric	552822	6480040	6550744	7002658	K=2
Purity	0.8550	0.855085	0.855085	0.855085	K=4
Entropy	0.9933	1.582708	1.84911	2.218392	K=5
NMI	0.9958	0.999014	0.983449	0.983945	K=2

Table V is used for determining the quality of clusters using external indexes. If Rand index is chosen as an external validity index, it forms the qualitative cluster at k=2(at which the Rand index is more for k value). If Adjusted rand index is chosen as an external validity index, it forms the qualitative cluster at k=4(at which the Adjusted rand index is more for k value). If Jaccard index is chosen as an external validity index, it forms the qualitative cluster at k=4(at which the Jaccard index is more for k value).
If Fowlkes mallows index is chosen as an external validity index, it forms the qualitative cluster at k=2(at which the Fowlkes mallows index is more for k value). If Rand index is chosen as an external validity index, it forms the qualitative cluster at k=2(at which the Rand index is more for k value). If Merkinn metric index is chosen as an external validity index, it forms the qualitative cluster at k=2(at which the Merkinn metric index is less for k value). If Purity index is chosen as an external validity index, it forms the qualitative cluster at k=4(at which the Purity index is more for k value). If Entropy index is chosen as an external validity index, it forms the qualitative cluster at k=5(at which the Entropy index is more for k value). If NMI index is chosen as an external validity index, it forms the qualitative cluster at k=2(at which the NMI index is less for k value).

VIII. CONCLUSION

This paper presents an analysis on churn prediction in telecom sector by using K-Means clustering algorithm. For a telecom churn dataset, chosen (k=2) to construct a cluster. Feature selection using information gain concludestotal day minutes, total day charge, customer service calls, international plan, voice main plan; total international minutes as important attributes. Customers those who talks more minutes in a day, doing more customer service calls or subscribed for voice mail plans and international calls, will not churn from the service. This ends with the K-Means clustering technique that best suitable for a given churn detection application with the validation criteria. There is a lot of challenging issues like handling clusters, SSE minimization and feature selection which are handled by K-Means algorithm. From this analysis, business organizations can improve their marketing strategies which indulge in profit maximization.

ACKNOWLEDGMENT

This work was supported by Laboratory of Computer science and engineering department, Gudlavalleru Engineering College, Gudlavalleru.

REFERENCES

1. Min chen, Simone A. Ludwig and Keqin li," Clustering in Big data", CRCPress, United states, 2017.
2. AdilFahad, NalaaAlshatri, ZahirTari, Abdullah Alamri, Ibrahi, "A survey of clustering Algorithms for big data: Taxonomy and Empirical Analysis", IEEE transactions on Emerging Topics in computing, Vol 2,Issue 3,pp.267-279, September 2014.
3. AvitaKatal, Mohammad Wazid, and RH Goudar "Big data: Issues, challenges, tools and good practices"In Contemporary Computing (IC3), 2013 Sixth International Conference on, pages 404{409. IEEE, 2013.
4. Wai-Ho Au, Chan, K. C. C., Wong, A. K. C., & Yang Wang. (2005). "Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data". IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2(2), 83–101.
5. Peng Li, Tingting Bi, Yang Liu, &Siben Li. (2014). "Telecom Customer Churn Prediction Method Based on Cluster Stratified Sampling Logistic Regression" International Conference on Software Intelligence Technologies and Applications & International Conference on Frontiers of Internet of Things 2014.
6. Chen, Y., Li, B., & Ge, X. (2011). "Study on Predictive Model of Customer Churn of Mobile Telecommunication Company". 2011 Fourth International Conference on Business Intelligence and Financial Engineering.
7. Aggarwal C, Zhai C. "A survey of text clustering algorithms. Mining Text Data" New York, NY, USA. Springer-Verlag: 2012. p. 77–128.
8. Huang Z. "A fast clustering algorithm to cluster very large categorical data sets in data mining"Proceedings SIGMOD Workshop Res Issues Data Mining Knowl Discovery.
9. EréndiraRendón, Itzel Abundez, AlejandraArizmendi and Elvia M. Quiroz "Internal versus External cluster validation indexes"; Issue 1, Volume 5, 2011.
10. Yun, C., Shin, D., Jo, H., Yang, J., & Kim, S. (2007) "An Experimental Study on Feature Subset Selection Methods"7th IEEE International Conference on Computer and Information Technology.
11. Chen, Y., Li, B., & Ge, X. (2011) "Study on Predictive Model of Customer Churn of Mobile Telecommunication Company"2011 Fourth International Conference on Business Intelligence and Financial Engineering.
12. Zerabi, S., &Meshoul, S. (2017) "External clustering validation in big data context"2017 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech).

AUTHORS PROFILE

V Renuka Devi M.Tech student in Gudlavalleru Engineering College. Graduated B.Tech from Gudlavalleru Engineering College. Gudlavalleru, (Andhra Pradesh). Her area of interest is Data mining, Big Data

Mrs.G.Bharathi received her Master degree in computer science and engineering from Gudlavalleru engineering college. She has been working as an assistant professor in Gudlavalleru Engineering College (Andhra Pradesh), Gudlavalleru. She is currently pursuing Ph.D. degree from JNTU-Kakinada. Her area of interest is Data Mining, Cloud Computing, and Big Data.

Dr. GUTTA V.S.N.R.V. PRASAD did his Ph.D.(CSE), MS Software Engineering in BITS Pilani and M.Tech in Computer Science and Technology in Andhra University. He is a member in various Professional Bodies. Presently working as Professor of CSE and Vice principal of academics at Gudlavalleru Engineering College, Gudlavalleru, (Andhra Pradesh).His area of interest is Data mining, Big Data, Network Security and Image Processing.