Hybrid Model for Pattern Discovery in Data Communication to Enhance Customer Relationship Management using Data Mining Techniques

B Arivazhagan¹ and DR R Sankara Subramanian²
¹Research Scholar, Department of Computer Science, Erode Arts and Science College, Erode, TN, India.
²Associate Professor, Department of Computer Science, Erode Arts and Science College, Erode, TN, India.

¹arivazhaganphd2019@gmail.com, ²rsankarprofessor@gmail.com

Abstract – Data Communication in customer oriented sector has significant information which is considered as a vital part in enhancing and improving customer base. Loyal customers in business sector are based on buying and selling of products or making use of services provided. Generally, the customers may switch on to variant enterpriser who affords best services with best price hence retaining them is a tedious task unless their communication data regarding service is analysed in periodic and to keep track of customer, the enterpriser should know the purchasing pattern and needs of loyal customers. The discovered pattern is also helps to afford proper discounts to the customers at the right time. This research paper experiments customer data in Telecommunication sector to discover useful and interesting patterns. There are so many data mining techniques put forward in revealing the pattern and this paper discusses all the possible mining techniques and evaluates an association rule mining with clustered data to create best rules. For clustering, Hierarchical agglomerative is used and FP-Growth algorithm is used for association rule mining. Both clustered data and association rules results regarding data communication are presented. The research work is implemented in weka tool and assessed with suitable evaluation metrics.

1. Introduction

Customer Relationship Management (CRM) [1] implies the procedures and goals that a sector follows while analyzing the consumer data. The entire relationship includes direct sales and service-related processes, analysis of customer trends and behaviors and predicting trust purchaser based on communication. There are some CRM [2] software that accumulates information in repository to provide analyst easy access for analyzing contacts, purchasing history and service. This helps employees communicate with clients, make updates, track performance and anticipate customer needs. The main purpose is to make interactions more efficient and productive. Automated Data mining procedure interact with the database and find the hidden knowledge about the customer in a large database by analyzing the data from different perspectives and summarizing it into useful information. The implied goal behind is to use the derived patterns in improving business practices.

1.1 Data Mining Techniques

Data mining explores patterns and new knowledge from the [3] collected data with numerous techniques which then can be used for various applications. It can be used in multi-disciplinary area such as Artificial intelligence, Machine learning with Statistics, Signal processing, Spatial & temporal data analysis, Economics, Business, Bio-informatics etc. The following are the data mining techniques used for all types of analysis,
1.1.1 Classification – Assemble various attributes together and split the instances into various categories, which are used to draw further conclusions. It contains class label that describes the instances with the attributes. The classification techniques differ based on statistical method used such as tree based, probability based, searching neighborhood based, neural network etc. but all techniques can be evaluated with same metrics. For example, the credit card holder can be classified based on the category ‘Low, Medium, and High based on the risk level.

1.1.2 Clustering - Groups data together based on the similarity with various using distance or similarity functions. Some distance functions includes euclidean, manhattan, minkowski and similarity functions cosine, jaccard, etc. For example, the customers can be grouped based on the usage of services. Mostly used existing clustering techniques are partition based, hierarchical based, density based. Each one differs with evaluation metrics also.

1.1.3 Association - Attributes entries that have high [4] correlation with another attribute entries are listed as best rules based on frequent itemsets. These frequent itemseta are generated with support values and rules are derived and ranked with support, confidence, lift, leverage and conviction etc. For example, if there is a link while a customer buys a specific item (Bread); there is a higher chance to buy a second item (Jam) and this rule is framed based on the previous purcjases or services included by the customer.

1.1.4 Outlier detection - In some cases, some instances or attributes doesn't have any similarity or it is far away from the rest then it is termed as outliers. For example, if a particular service is not consumed by the customer for a long time, then it ids considered as outliers from the rest and it should be found out to better understand about the target audience in the process. Some outlier techniques are local outlier factor, class outlier factor etc. The outlier can also be identified with classification and clustering with some extra procedures.

1.1.5 Prediction - Recognize and understand the historical trends to chart accurate forecasting the events happen in upcoming period. For instance, review of consumer’s credit histories and past transactions will help to foretell about the buying behavior of consumers in future. Regression techniques such as linear regression which are used for real data type and logistic regression which are used for binary data types for prediction.

2. Literature review

Anika Singh et al [5] proposed a model to segment customers with K-Means and Hierarchical clustering techniques. Market data is taken for analysis to know the expectation of customers and group the customers based on their preferences. For clustering, distance metric is the key component as each metric gives different clusters. Data are standardized within scale in prior. K-Means require number of clusters as its primary input while hierarchical doesn’t need. The hierarchical model in this paper is cut into clusters after the dendrogram is produced. With the elbow method the number of cluster is selected as three for hierarchical method. Both clustering uses Euclidean distance as the metric. But K-Means produce different clusters for each time as it fully depends on the centroid.

Shreya Tripathi et al [6] explored the importance of customer segmentation in CRM through Centroid based clustering and Hierarchical based clustering. Target marketing is completely depends on grouping customers based on their purchasing behaviors and needs. In Clustering, data points are clubbed together in which same cluster has more similarities. They are internally homogeneous and externally heterogeneous. Centroid based clustering K-Means (with five clusters) and Hierarchical clustering use Euclidean distance to form clusters. Number of clusters is not prefixed in Hierarchical and it is produced huge number of clusters which is not able to visualize. Hence a cut-off line is used in the dendrogram for five clusters to make visualization better. The quality of hierarchical clustering improves when compared to K-Means clustering with the increased number of K Values.

Phan Duy Hung et al [7] implemented hierarchical clustering algorithm for the segmentation of customers on credit card to focus on the marketing ideas. The aim is to segment customer by dividing
the user on to specific group to provide offers. The credit card dataset summarizes the usage of active card holders within six months and it includes eighteen features where the credit limits, purchase and charges were used for segmentation. Initially, missing values and outliers are handled to get a structured dataset. Agglomerative hierarchical clustering is performed to get sub groups of customers. To cut the dendrogram into meaningful clusters, three types of methods namely Elbow, Silhouette and Gap Statistic methods are used and based on this method three clusters are preferred. These three clusters define the variant customers based on the selected attributes.

Mehmet Ali Alan et al [8] proposed association rule to explore the pattern between sold products using sales data of a supermarket. Association rule mining method Apriori is used to reveal the purchasing behavior pattern of the customer. In this paper, sales obtained from a supermarket in the city centre were used as data. Two criteria “support” and “confidence” are used in order to reveal the relations between the products. Tanagra software was used in the study and association rule mining Apriori was performed between the product sales made to 2205 customers.

Adebola Orogun et al [9] develop an association rule to predict customer using online store and derive interesting patterns from customer purchasing. Online (E-Commerce) data is taken from UCI repository. Initially preprocessing method is carried with Pearson correlation to select attributes then association rule mining is implemented. Performance is assessed with min support with number of frequent items and confidence with number of association rules generated. Least number of support and confidence gives more rules and frequent item sets.

Narasingha Rao et al [10] compare three association rule algorithms namely Apriori, FP-Growth and Eclat. The data of an online shopping website is tested for customer behavior. The analysis helps to assess customer that lead to improve the quality of service. From the study, it is revealed Apriori algorithm takes more scan to generate item sets where as FP-Growth takes less scan and Eclat can be utilized efficiently for small datasets. But finally, it was found that Apriori is more suitable for obtaining customer in online shopping.

3. Methodology

Telecom dataset from Kaggle repository is taken for analysis regarding data communication in the proposed method. The dataset comprise of the fields namely Customer Id, detail of the customer whether having partner, dependents, tenure in years, telephone service, Multiple lines, Net service, having security, backup, having device security, Technical support details, Streaming TV and Movies, Contact information, about Billing, Pay mode details, Charges per month and Total charge. The telecom service provider is offering six services. Initially, in the proposed method ARHAC-TELE the dataset is clustered using hierarchical agglomerative clustering with inner product similarity and the clustered set is applied for association rule mining to derive best rules. The clustered set has the identity of group and from each group best pattern is derived instead of deriving from the whole set hence the model is optimized to enhance customer relationship management.

Proposed method – Association rule with Hierarchical agglomerative clustering for Telecommunication sector (ARHAC-TELE)

3.1 Hierarchical agglomerative complete linkage clustering with inner product similarity

It is one of the methods in cluster analysis that produces hierarchy of clusters with the objects. There are two types namely Agglomerative and Divisive in Hierarchical clustering. Agglomerative is a Bottom – Up approach in which initially each observation begins from its own cluster and merged based on distance function iteratively until all observation belongs to same cluster and divisive is Top-Down approach.
Complete Linkage - Two clusters are merged if the maximum distance between the points \((b,c)\) in the clusters are less than or equal to the threshold distance.

\[
\text{Complete Linkage} = \max \{d(b,c)\} \tag{1}
\]

Inner product similarity - It changes the vector space with an additional structure by taking product (inner product) of every couple of vectors in scalar quantity. This product denotes length of vector or angle between the vectors. It is a generalization of Euclidean space by taking scalar dimension.

For the Euclidean space given as \(<(x_1, x_2, \ldots, x_n)>, <(y_1, y_2, \ldots, y_n)>\)

\[
\text{Inner Product} = x_1 y_1 + x_2 y_2 + \ldots + x_n y_n \tag{2}
\]

Generally, it satisfies the properties such as Linearity, Symmetric property and Positive Definite Property.

### 3.2 Association rule with FP-Growth

It is a machine learning method that consists of rules to discover interesting relation within features in huge database with measures. The well used algorithms are FP-Growth and Apriori but they mine frequent itemsets. Additional steps are used after mining frequent items to generate rules.

The rule is defined as,

\[
B \Rightarrow A, \text{in which } B, A \subseteq J \tag{3}
\]

Where, \(J=\{j_1, j_2, \ldots, j_m\}\) be the set of ‘m’ attributes known as itemsets generated from FP-Growth and \(N=\{n_1, n_2, \ldots, n_m\}\) be the set of \(m\) transactions known as database. Every rule is composed by two different items known as antecedent (B) and consequent (A) and the rule is composed as \(B \Rightarrow A\).

### 3.3 Process in FP-Growth

The algorithm is carried out in two passes. In the initial pass, it counts the existence of items and caches these counts in a 'header table'. In the upcoming passes, the algorithm creates the FP-tree by including transactions into a prefix tree. These are sorted in descending order based on the frequency. The tree growth starts from bottom of the header table. Nodes with minimum support are pruned and the resulting nodes are the frequent item sets then the association rule creation begins.

### 3.4 Process in generating Association rule

Association rules are formed with generated frequent item sets with preprocessing. It consists of two constraints namely support and confidence which are user defined parameters. It comprise of two steps as follows:

1) min support - find frequent itemsets in a database.
2) min confidence - generates frequent itemsets in order to form rules.

### 3.5 Measures used to derive rules

1) **Support** - States the frequency of items in the dataset. ‘T’ is the set of transactions and ‘t’ is the proportion of actions which consists the itemsets A.

\[
\text{Support}(A) = \frac{|\{t \in T; A \subseteq t\}|}{|T|} \tag{4}
\]

2) **Lift** - The ratio of observed support with expected support

\[
\text{Lift}(C \Rightarrow D) = \frac{\text{Support}(C \cup D)}{\text{Support}(C) \times \text{Support}(D)} \tag{5}
\]
3.6 Advantages

- Inner product similarity is used instead of Euclidean distance to reduce computational complexity and hence increase the accuracy.
- Complete Linkage criteria in hierarchical clustering forms groups with maximum distance leads to optimum clusters.
- Association rule mining with the clustered set (hybrid model) eases the frequent itemset generation and finds best rules or patterns from each group.

4. Results and Discussion

The results are assessed in weka tool with the evaluation metrics namely number of itemsets, number of association rules.

4.1 Screen in weka

The experimental screens in weka with clustered instances, clustered charts, generated rules are depicted in Figure 1, 2 and 3.

![Clustered Instances](image1)

Figure 1. Clustered instances in weka

Figure 1, shows the distributed instances with five clusters. The number five is chosen to have less intra and high inter cluster distances. Cluster 0 has 893 instances, Cluster 1 has 390 instances, Cluster 2 has 1016 instances, Cluster 3 has 1074 instances and Cluster 4 left with 535 instances. All instances are grouped based on the similarity in data. These clustered instances is labeled with cluster number and applied for association rule mining.

4.2 Association rules with clustered set

![Association rules](image2)

Figure 2. Association rules
Figure 2. shows the sample rules from clustered set. Other than the ordinary attributes in the dataset, cluster ID is also included with the default measures confidence, lift, leverage and conviction. In this research work, user minimum support value is used to generate frequent itemsets and lift is used as a primary factor to form rules. Hence in this screen shot, lift value is shown in decreasing order where the higher value of lift value produce most frequent pattern and the lower value depict the least frequent pattern. Other measures are showed based on the lift value.

5. Performance analysis

The performance is assessed with number of rules generated, number of itemsets generated, number of cycles performed for variant support and lift values for existing Association rule mining for telecommunication (AR-TELE) and proposed Association rule mining with Hierarchical agglomerative clustering for telecommunication (ARHAC-TELE).

Table 1. Performance analysis

| Minimum support | Lift | Number of cycles performed | Number of itemsets | Number of rules |
|-----------------|------|-----------------------------|--------------------|-----------------|
|                 |      | AR-TELE | ARHAC-TELE | AR-TELE | ARHAC-TELE | AR-TELE | ARHAC-TELE |
| 0.01            | 0.1  | 20      | 18         | 386     | 727        | 1000     | 2500        |
| 0.03            | 0.3  | 20      | 18         | 137     | 289        | 584      | 1810        |
| 0.05            | 0.5  | 19      | 16         | 85      | 177        | 292      | 990         |
| 0.07            | 0.7  | 19      | 16         | 58      | 120        | 186      | 488         |
| 0.09            | 0.9  | 18      | 15         | 44      | 80         | 86       | 240         |

From Table 1, three types of evaluation metrics are discussed with variant combination of support values to generate itemsets starts from 0.01 to 0.09 and lift values to form rules starts from 0.1 to 0.9. With the combination of support and lift values the proposed method ARHAC-TELE generates more frequent itemsets and more association rules than the existing AR-TELE.

Also it is noted,
- Number of rules highly depends on number of itemsets and for more itemsets, more rules are formed.
- More number of itemsets and rules are formed for least value of support and lift.
- Number of cycles performed to generate rules is less for proposed ARHAC-TELE.
- Number of itemsets and rules are generated in high number for proposed ARHAC-TELE.
Figure 3. Number of cycles in performing rule generation

Figure 3, shows number of cycles performed for existing AR-TELE and proposed ARHAC-TELE. The hybrid model ARHAC-TELE with clustered instances takes less cycles in generating rules.

![Number of frequent itemsets](image)

Figure 4. Number of frequent itemsets

Figure 4, shows number of frequent itemsets for AR-TELE and ARHAC-TELE. The proposed model ARHAC-TELE for telecommunication generates more frequent itemsets with clustered data as it has one additional labeled attribute with cluster ID.

![Number of Association rules](image)

Figure 5. Number of rules

Figure 5, shows number of rules for AR-TELE and ARHAC-TELE. The proposed ARHAC-TELE forms more association rules as it has more frequent itemsets.

6. Conclusion and Future work

Data communication in Tele communication sector has significant pattern and it should be analyzed to promote the growth and to develop business strategies. This research work applies data mining techniques clustering and association rule to develop a hybrid model to reveal the pattern. The hybrid model use Hierarchical agglomerative clustering and FP-Growth association rule mining to generate itemsets and rules. The itemsets are generated with minimum support value and rules are formed with
lift values and the work is evaluated in WEKA tool. From the analysis it is noted that the proposed model ARHAC-TELE takes less cycles to perform and produce more itemsets and rules.

The work can be tried with other data communication sectors and variant clustering techniques with other distance or simialiry functions can be applied.

References

[1] Loudon, D.L., 1993, Customer Behaviour: Concepts and Applications. McGraw-Hill, New York.

[2] Michael, J.A.B. and Gordon, 2011, S.L. Data Mining Techniques: For Marketing and Sales, and Customer Relationship Management. 3rd Ed., Wiley Publishing Inc., Canada.

[3] Jiawei Han and Michelin Kamber, August 2000, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, ISBN 1-55860-489-8.

[4] Pavels Goncarovs, December 2017, Data Analytics in CRM Processes: A Literature Review, Information Technology and Management Science, vol. 20, pp. 103-108.

[5] Anika Singh, Prasanna Balaji Venkataraman, 2017, Customer segmentation through K-Means and Hierarchical Clustering Techniques, IJSRD - International Journal for Scientific Research & Development| Vol. 5, Issue 07, pp.512-516.

[6] Shreya Tripathi, Aditya Bhardwaj, Poovammal, , 2018, Approaches to Clustering in Customer Segmentation, International journal of Engineering and Technology, Vol.7(3.12), pp. 802-807.

[7] Phan Duy Hung, Nguyen Thi Thuy Lien, Nguyen Duc Ngoc,2019, Customer Segmentation Using Hierarchical Agglomerative Clustering, ICISS, Association for Computing Machinery (ACM), pp. 33-37.

[8] Mehmet Ali Alan, Ali Riza Ince, (2016 ), Use of Association Rule Mining within the Framework of a Customer-Oriented Approach, European Scientific Journal, Vol.12, No.9, pp. 81-99.

[9] Adelboa Orogun, Bukola Onyekwelu, (2019), Predicting Consumer Behaviour in Digital Market: A Machine Learning Approach, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 8, Issue 8, pp.8391-8402.

[10] Narasingha Rao, Sita Ratnam, Prasanth, Lakshmi Bhavani, 2018 , A Survey of online consumer Behavior using Association Rules, International Journal of Engineering & Technology,Vol 7(2.32), pp.206-208.