Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Analytics Model for TelecoVAS Customers' Basket Clustering using Ensemble Learning Approach

Mohammadsadegh Vahidi Farashah
Islamic Azad University Khorasgan Branch
Akbar Etebarian (etebarian@khuisf.ac.ir)
Islamic Azad University of Khorasgan https://orcid.org/0000-0002-2330-8521
Reza Azmi
Alzahra University
Reza Ebrahimzadeh Dastjerdi
Islamic Azad University Khorasgan Branch

Research

Keywords: Basket Analysis, Value Added Service, Ensemble learning

Posted Date: November 18th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-107395/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published on February 17th, 2021. See the published version at https://doi.org/10.1186/s40537-021-00421-1.

An Analytics Model for TelecoVAS Customers' Basket Clustering using Ensemble Learning Approach

Mohammadsadegh Vahidi Farashah[a], Akbar Etebarian[b*], Reza Azmi[c], Reza Ebrahimzadeh Dastjerdi[d]

[a]Mohammadsadegh Vahidi Farashah

Ph.D. Candidate of Information Technology Management, Department of Management, Isfahan (Khorasgan) Branch, Islamic Azad University, Isfahan, Iran.

sadegh.vahidi@khuisf.ac.ir, +989125055404

[b*]Akbar Etebarian

Associate Professor of Management, Department of Management, Isfahan (Khorasgan) Branch, Islamic Azad University, Isfahan, Iran. Corresponding Author: etebarian@khuisf.ac.ir, +989132105784

[c]Reza Azmi

Associate Professor of Computer Engineering, Department of Engineering and Technology, Alzahra University, Tehran, Iran.

azmi@alzahra.ac.ir, +989121547174

[d]Reza Ebrahimzadeh Dastjerdi

Assistant professor of Management, Department of Management, Isfahan (Khorasgan) Branch, Islamic Azad University, Isfahan, Iran.

Ebrahimzadeh@gmail.com, +989132252272

An Analytics Model for TelecoVAS Customers' Basket Clustering using Ensemble Learning Approach

Abstract

Value Added Services at Mobile Communications Company provide customers with a variety of services. Value added services generate significant revenue annually for telecommunications companies. Providing solutions that can provide customers of a communications company with relevant and engaging services has become a major challenge in this field. Numerous methods have been proposed so far to analyze customers' carts and provide related services. Despite the many applications that these methods have, they still face difficulties in improving the accuracy of bids. This paper combines the X-Means algorithm, the ensemble learning system, and the N-List structure to analyze the customer portfolio of a mobile communications company and provide value-added services. The X-Means algorithm is used to determine the optimal number of clusters and clustering of customers in a mobile communications company. The ensemble learning algorithm is also used to assign categories to new Elder customers, and finally to the N-List structure for customer basket analysis. By simulating the proposed method and comparing it with other methods including KNN, SVM, and deep neural networks, it has improved the accuracy of about 7%.

Keywords: Basket Analysis, Value Added Service, Ensemble learning.

Introduction

Value Added Services One of the important features and capabilities of mobile communication companies is that it enables customers to receive services and services by paying to mobile communication companies. These services can be very useful and effective in analyzing customer behavior [1,2]. Customer Behavior [3]. In many online systems, processes are referred to as the customer doing on a continuous basis. These

operations and transactions can be repeated in a few days [4]. Customer basket analysis is one of the most widely used data mining methods to analyze the goods in one or more baskets that the customer analyzes at a particular moment [5]. The basket analysis program can be designed and run in a supermarket not only because of the ability to help with sales promotional design but also because of the ability to become a reference for re-managing items in stock [6].

In recent years, customer-generated transactions are commonly used as information for analysis. This article also reviews or re-examines customer transactions to gain valuable information. For example, information about an item that sells higher. In addition, information can be used to add stocks to this sample. Also, these transactions and customer performance can be used as the equation of each item purchased in the customer basket. Using this information, it can be used to display the right product to attract customers. One of the most important uses of these transactions is data analysis and transaction and customer basket [7].

Customer basket analysis is one of the modes of analysis based on customer behavior. Whereas shopping in the supermarket is through the identification and direct linkage between different items with the customer [2]. With regard to analyzing customer baskets as well as identifying items that are often purchased by them, there are challenges today that can be attributed to not recognizing customer behavior, product groups that have the most repeat purchases, product alignment to increase Sales pointed out. Using the customer basket analysis approach, we can identify items that are often purchased by customers at the same time and provide an opportunity to enhance the performance of the telecommunications value added service system.

There are challenges and difficulties in how to provide services in value-added telecommunication systems such as inadequate accuracy and high error of providing

related services to the customers. Until now, there are various methods for analyzing customers' portfolio such as the method of customer basket analysis based on their transaction records [8], customer basket analysis approach by process category [9], portfolio analysis approach. Customer Acquisition with Apriori Algorithm [10], Customer Basket Analysis Approach Using a Combination of Artificial Intelligence Techniques and Associated Laws and Minimal Spanning Tree [11], Customer Basket Analysis Approach with the Advance System Business Strategy Forecast [12], Improving the approach of customer basket analysis in an efficient way called feasibility Utility Mining [13], is provided.

Most of the approaches presented have challenges and problems such as inadequate consideration of metric and factors related to customer behavior, inadequate quality of services provided to specific and related customers, inaccuracies in macro data analysis and so on. [11]. In this paper, we use the N-List algorithm-based technique to analyze the customer basket and increase the accuracy of customer basket analysis using the proposed ensemble learning system. The proposed N-List algorithm ensures that the comprehensiveness is maintained and the service execution speed is increased. The proposed ensemble learning system in this research consists of combining three machine learning algorithms including deep neural networks, C4.5 decision tree and SVM-Lib algorithm. The proposed ensemble learning system is based on maximum votes and sends the best response to the output at each step. The remainder of this paper is divided as follows: Section 2 reviews the work done in the past, Section 3 describes the proposed approach and architecture. In Sections 4 and 5 the results are obtained and the final conclusions are discussed.

Related work

In 2019, Jiang et al proposed a new methodology for dynamic modeling of customer preferences on products based on their online reviews, which mainly focused in mining ideas from online reviews and using customer preferences to develop dynamic model by using DENFIS approach. Unlike the conventional DENFIS approach which only provides crispy outputs in its modeling, the proposed DENFIS approach is capable of providing fuzzy outputs as well as crispy ones. By predicting fuzzy outputs, companies can face to the worst-case and the best-case scenario of customer preferences while designing their new products, services [14].

In 2018, Musalem and his colleagues presented a customer basket analysis model based on process categories. The basis of their work in this study was based on the similarity and distance between the existing samples, one of the most important benefits of their work being the speed of analysis of the customer's basket. One of the major disadvantages of this model is the lack of proper accuracy for online portfolio analysis, the lack of comprehensiveness and the fact that the model does not perform well on large data sets. The performance range of the methodology proposed in this study is at the supermarket level and has a poor performance for the larger statistical population [9]. In 2018, Szymkowiak and his colleagues proposed an Apriori algorithm for customer basket analysis. The Apriori associative algorithm has an infinite constraint on the large statistical population. In their research, they have been able to apply the data and items of a supermarket to achieve the desired accuracy. Therefore, one of the most important advantages of this model is that it has a good basket analysis speed, medium accuracy and one of the major disadvantages of this research is its lack of comprehensiveness and high flexibility [10].

In 2018, Jain and his colleagues presented a customer basket analysis model with the help of a business strategy forecasting system. They carried out the process of analyzing

the customer basket based on business logic and statistical business. They used statistical methods to make the service closer to the person concerned. One of the most important advantages of the method presented in this article is the accuracy of the service provided to the customers. In addition, the implementation time of the method proposed in this study was moderate but not comprehensible for large and large spaces [12].

In 2018, Srivastava and his colleagues used a portfolio optimization model of customer shopping in an efficient way called mining. They proposed an improved mode of data mining called utility mining. With the help of the technique provided, they were able to quickly and accurately perform the customer basket analysis process, but they did not have the potential and high development potential [13]. In 2017, Kurniawan et al. Presented a customer basket analysis model based on their transaction records. In their research they used associative and data mining techniques such as neural networks and Apriori. One of the most important benefits of their work is the speed of analysis of the customer basket. One of the major disadvantages of this model was the lack of precision for online portfolio analysis, the lack of comprehensiveness and the fact that the model does not perform well on large data sets [8]. In 2016, Kaur and his colleagues proposed a customer basket analysis model using a combination of data mining methods and association rules. In their methodology, they used data mining to improve the accuracy of customer basket analysis. They have also used data mining techniques such as neural networks and other machine learning techniques to teach based on purchase information and customer transactions. One of the most important advantages of their method is having sufficient accuracy in analyzing and analyzing the customer cart. Their analysis is very slow and their production model is complex. It does not support a large statistical community and operates within the supermarket. The method proposed

by them does not have the potential for future growth [5]. In 2016, Venkatachari and his colleagues used a combination of associative approaches such as Apriori and FP-Growth to analyze customer baskets. Their proposed strategy is based on sharing repeated transactions. One of the benefits of their approach has been to improve the accuracy and consistency of customer basket analysis. One of the major disadvantages of their method is the increased runtime and lack of potential for development in the larger statistical community [15]. In 2015, Sherly and his colleagues used parallel and distributed techniques and associative rules to analyze the customer basket. In their research, they sought to increase the speed and completeness and accuracy of customer basket analysis. Eventually they succeeded in increasing the accuracy to some extent and improving speed and comprehensibility significantly with parallelization [16]. From the analysis of the research that has been done so far, it can be seen that many of the researches suffer from inadequate accuracy, speed of cart analysis, inadequacy and so on. Thus, despite such problems in the models proposed in the context of value-added customer basket analysis, this paper presents a process-based approach and algorithm for extracting iterative patterns such as N-List. The value proposition system proposed in this article increases the accuracy of customer basket extraction and analysis. The process approach with the help of deep neural networks algorithms, C4.5 decision tree and SVM-Lib algorithm significantly enhances the quality of value-added services provided.

The proposed method

The proposed method in this paper is based on X-Means clustering algorithms, N-List structure for extracting frequent patterns, and ensemble learning system to provide attractive value added services to telecommunication customers. This section describes

the stages of service delivery using the proposed hybrid approach. Important parts of the proposed method are:

Data normalization

Data normalization is used to increase clustering accuracy. At the preprocessing stage, in order to obtain better results, we normalize the behavior information of the telecommunications customers between [0,1]. In other words, all datasets are mapped into matrices, and matrix rows are normalized. Normalization is due to higher accuracy. To normalize the values of each dataset, we use (1).

$$\text{Normalize}(x) = \frac{(x - X_{min})}{(X_{max} - X_{min})} \quad (1)$$

Where Xmax and Xmin are the maximum and minimum values in the range of my X property. After normalizing the data, the values of all the attributes fall within the range [0,1].

Customers Clustering using XK-Means Algorithm

In this paper, we user hybrid of K-Means and X-Means algorithm together for clustering customer based on behavior information. Combine K-Means and X-Means clustering algorithm called XK-Means algorithm.

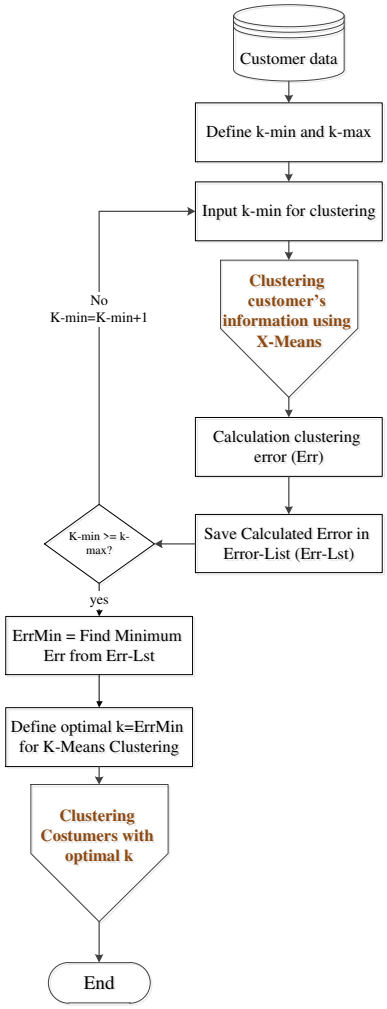


Figure 1. Flowchart of X-Means Clustering Algorithm for clustering customer's information.

The second phase of this paper id of customers clustering. Customers may have two cases. One case a new customer that is activity in the system. Another case is that the intended customer is already registered in the system and that there is activity in the system. The X-Means clustering algorithm receives behavior information of customers as. It then moves the customer to a cluster based on behavior information. The X-Means algorithm is used to cluster the customer's information. One of the basic applications of using the X-Means clustering algorithm in the proposed method is to apply cluster (labels) on customer's information that are unattended and do not have label properties.

The Fig.1, illustrates the application of the X-Means clustering algorithm to the clustering of each customer's information.

As can be seen from Fig. 1, all customers of the telecommunication company were first introduced to the X-Means algorithm in order to calculate the optimal K value using this algorithm. The X-Means algorithm runs in the background on powerful telecommunication servers. Because the X-Means algorithm is slow and has a high time complexity. After determining the number of optimal clusters (*K*), the K-Means algorithm with the optimal *K* number is used for clustering.

The K-Means algorithm is a basic clustering algorithm that performs the clustering process of samples based on a number of clusters called *k*. One of the most important disadvantage of the K-Means algorithm is that the number of clusters has to be determined by the researcher and based on this amount of clustering process. Determination of *k* was highly error-free and often did not provide optimal clustering. Unlike the K-Means algorithm, which has a high speed and receives a number of *k* from the input, this algorithm has a relatively low speed but instead obtains the optimal *k* number and yields the number of clusters with the lowest error rate as the cluster.

It uses this number of clusters as input to the K-Means algorithm and performs clustering of the customer's information. After customer's information is clustered, outlier's samples that behave similar to other samples are removed from the dataset. The K-Means algorithm steps is as follows:

1. Select the number of *k* for the number of clusters.
2. Then the *k* center for all data is randomly generated ($\mu_1,\ldots,\mu_k$).
3. Then repeat the following steps until the convergence is complete:

   Calculate c for each i

$$c^{(i)} := argmin_j \left|\left|x^{(i)} - \mu_j\right|\right|^2 \quad (2)$$

For each j, calculate the value of μ as follows and j is the value.

$$\mu_j := \frac{\sum_{j=1}^{m} 1\{c^{(i)}=j\}x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)}=j\}} \quad (3)$$

After the clustering operation is completed, all customers fall into their respective clusters. In Fig. 2, the internal structure of the X-Means algorithm is visible.

Figure 2. X-Means Clustering Algorithm [17].

Algorithm: Extended k-Means (E-km)(S,C)
Input: S: List of segments in MOD, C initialized k cluster centroids, ¥ Give Threshhold
Output: C_{List}: List of Clusters
foreach (s ϵ S) **do**
foreach (c ϵ C) **do**
TempDist=Direction Evaluation (s, direction, c.
direction) + EuclideanDistance (s, c);
end
MinDistance=Min[TempDist]. centroid;
ClosetCentroid=Min[TempDist]. centroid;
if (MinDistance ≤ ¥) **then**
Cluster=C[ClosetCentroid];
C_{List}= Update.Centroid(Cluster, s);
else
$Cluster_{New}$. centroid=s;
C. Add ($Cluster_{New}$. centroid);
end
return$C_{List;}$

As can be seen from Fig. 2, the initial k number is first determined. Then the K-Means clustering algorithm is repeated with the same number k. The error rate is calculated and then one unit is added to the number of clusters and the previous steps are executed again. This procedure will continue until the best value of k is calculated.

This paper uses the X-Means clustering technique, which is an extended version of K-Means, to assign labels to new customers. So the input of the X-Means algorithm is the customers of the telecommunications company. The output of this algorithm is k. Finally, the number k is applied to the K-Means clustering algorithm. The input of the

K-Means clustering algorithm is for customers of the telecommunications company and the output of this algorithm is labeling for customers. In the table 1, show sample of cluster and labeling customer's information.

Table 1. sample of cluster and labeling customer's information.

Customer ID	Age	Sex	Cluster
C1	25	0	Cluster_1
C2	30	1	Cluster_2
C3	45	1	Cluster_2
C4	32	0	Cluster_1
C5	29	1	Cluster_2
C6	51	0	Cluster_1
C7	22	0	Cluster_2

These clusters are used as a label for each customer. Each Ci is labeled after the customers of the telecommunications company are clustered using the XK-Means algorithm. Up to this point a set of customers clustered with specific tags is available. So, we used X-Means algorithm for finding optimal k for K-Means clustering algorithm.

The ensemble learning

In the Fig. 3, shown Ensemble learning flowchart for classify customers. Fig. 1, implementation in first rectangle in Fig. 2, for clustering customer's information using XK-Means. At the core of the ensemble learning are the most popular classification algorithms such as deep neural network, the C4.5 decision tree with the Information Gain kernel and the SVM-Lib algorithm for classification new customers in mobile telecommunications companies. New customers are categorized based on their behavior information. Category assignment for new entrants allows more accurate value-added services to be offered to customers based on services purchased by others. In the ensemble learning system, in-depth learning with 50 hidden layers, the C4.5 decision

tree is combined with the Information Gain core and the SVM-Lib algorithm, and at each stage the best batch is selected from the batches presented as the end result for New customer specified.

The training data, which is 70% of the data, is entered into the algorithms and the corresponding model is generated. Experimental data are also entered into the models produced to determine a category based on behavior information. Suppose Test 1, a male customer, aged 35 years, lives in X Province. This example is now entered into the Deep Learning Algorithm model and specifies the category 1 deep learning for Sample 1. Sample 1 also joins the decision tree algorithm and this algorithm specifies category 2 for sample 1. Finally, for example 1, the SVM-Lib algorithm defines batch 1. Outputs 1,2 and 1 are assigned to the Max system and based on the maximum votes, output 1 is determined for sample 1. In Category 1, for example, customers are between the ages of 30 and 40 and are male in X province. Thus the output of the ensemble learning system is as follows.

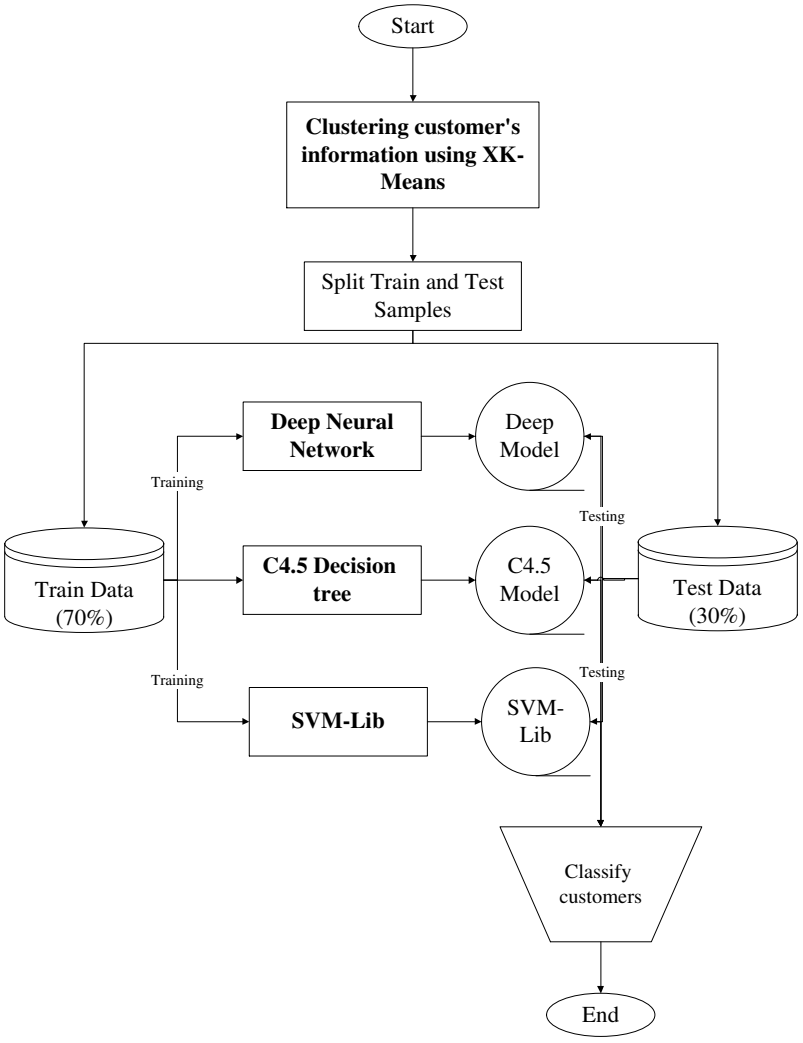


Figure 3. Ensemble learning flowchart for classify new customers.

In the ensemble phase, a new category is selected for new customers. Customers in the target group behave similarly to other Customers. After the process search system was implemented and the new category was assigned to the new customers, the N-List structure was implemented on all customer baskets in the selected category, and finally, based on the analysis, a set of services to New customers are provided.

Basket Analysis using N-List Algorithm

One of the most important steps in this paper is to analyze the basket of customers interested in receiving value-added services based on their behavior extraction and

customer transaction records in the telecommunications system. In this study, the N-List associative algorithm is used to analyze the customer cart [18]. Based on its tree structure, the N-List algorithm processes customer transactions and offers customer services based on extracted repetitive rules and transactions. Based on repetitive transactions, a set of features that are effective in repetitive transactions are extracted and then used in the ensemble learning system. Suppose a database called DB has n transactions and these transactions have a number of items. For example, the following table shows a sample DB dataset with 6 transactions (n = 6).

Table 2. A sample DB dataset with 6 transactions.

Transaction	Items
1	*A, C, T, W*
2	*C, D, W*
3	*A, C, T, W*
4	*A, C, D, W*
5	*A, C, D, T, W, E*
6	*C, D, T, E*

This small data set is used to illustrate how the basket is analyzed in the proposed system. The value of Sup for the X pattern is represented as σ (X), where X∈I and I are the set of all items in the DB dataset and the number of transactions that contain all the items in X. A pattern with a k-item number is called a k -pattern, and I1 is a set of duplicate patterns arranged in descending order. For convenience, the pattern {A, C, W} is written in ACW. Set the minSup (minimum threshold sup) to a certain threshold. Suppose the DB dataset in table (1) is minsup = 50%. AW and ACW are two of the most frequent patterns because δ (AW) = δ (ACW) = 4> 50%. Now with this prerequisite, the structure of the N-List algorithm to extract repetitive transactions is discussed.

In 2012, Deng and Xu introduced a tree structure called the PPC tree. In the PPC tree, each tree node has five values of n (N_i). f (N_i). child (N_i). pre (N_i). post (N_i)

[18]. The N-List algorithm or structure is based on the PPC tree. The N-List structure has a set of nodes. Each node in the N-List structure is represented as N_i. Each n_i node consists of a pp code. The pp code value of each N_i node in a PPC tree contains an instance of the form C_i = <pre (N_i). pre (N_i). f (N_i). The N-list associated with pattern A is represented as NL (A). A set of PP codes from PPC tree nodes associated with pattern A. The value of NL (A) of pattern A is calculated based on relation (4).

$$NL(A) = \bigcup_{\{N_i \in R \,|\, n(N_i)=A} C_i \tag{4}$$

Where C_i is the PHP code for N_i support for A. The value of δ (A) is calculated as follows:

$$\delta(A) = \sum_{C_i \in NL(A)} f(C_i) \tag{5}$$

In the above relation, the N-List is associated with k-patterns. Suppose XA and XB are two k-1 patterns with the prefix X (can be an empty set) such that A exists before B in order I_1. If XA and XB are two repetitive patterns (XA is a repetitive pattern before XB and X can be an empty set). Then NL (XA) and NL (B) are the N-lists associated with XA and XB, respectively. Given the N-list method with a NL (XB) ⊆NL (XA) k pattern:

$$NL(XAB) = \bigcup < pre(C_i).post(C_j).f(C_i) > \tag{6}$$

That $C_i \epsilon NL(XA)$ and $C_i \epsilon NL(XB)$ and C_i Parent C_j is.Therefore, $\sigma(XAB) = \sum C_i \epsilon NL(XAB) f(C_i) = \sum C_i \epsilon NL(XB) f(C_i) = \sigma(XB)$is.Figure (1) illustrates the creation of a PPC tree using the DB example with %minSup = 50.

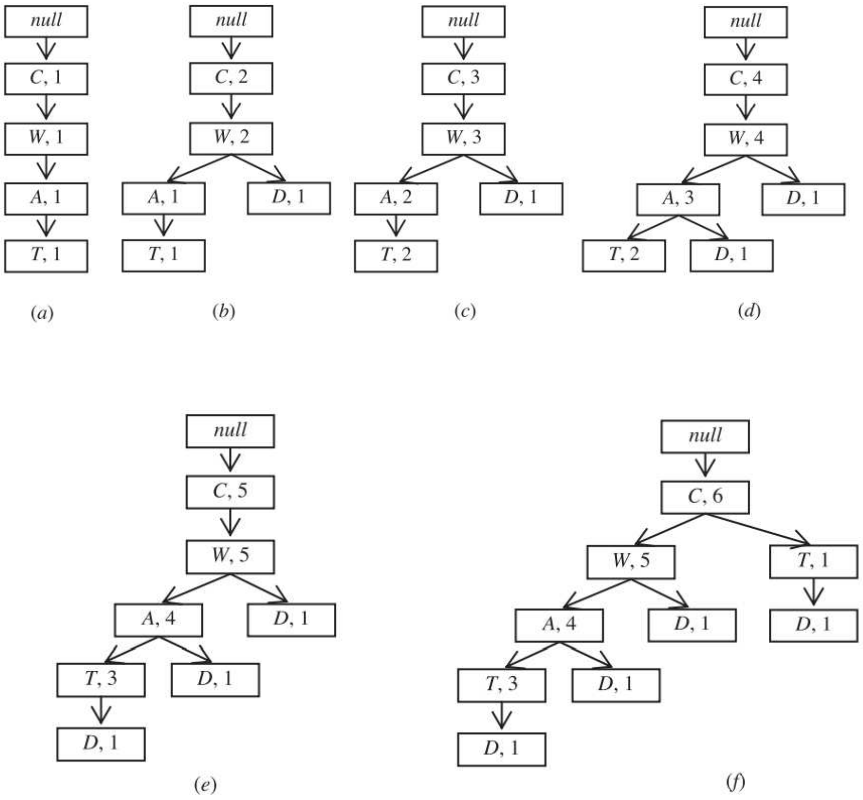


Figure 5. Illustration of PPC tree creation using DB example with minSup = 50%.

The N-List algorithm first creates the PPC tree and then generates it to generate N-lists associated with the repetitive sets 1. Then, the divide and conquer strategy is used to use PPC. In the following, for example, the N-List structure implementation process is described in order to find frequent patterns.

Consider the DB dataset example in Table. 3, with minSup = 50% to illustrate the performance of this algorithm. First, the N-List algorithm removes all items that do not meet the minSup threshold frequency and arranges the remaining items in descending order. The algorithm then, in turn, imports the remaining items in each transaction into the PPC tree.

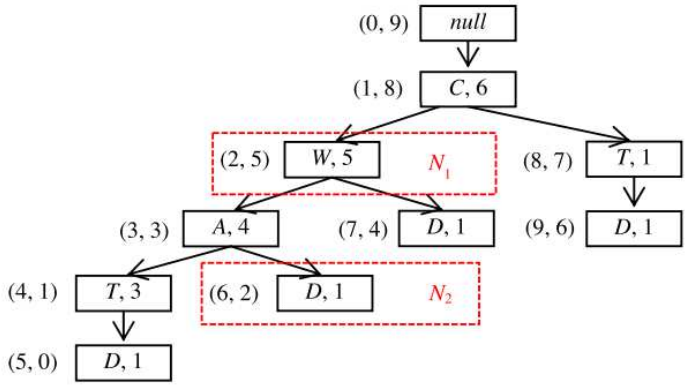


Figure 6. The final PPC tree created from the DB example with minSup = 50%.

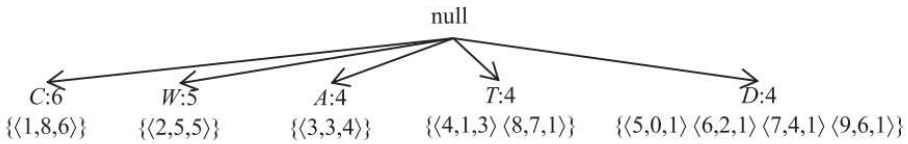


Figure 6. One-repetition frequency patterns and their N-List.

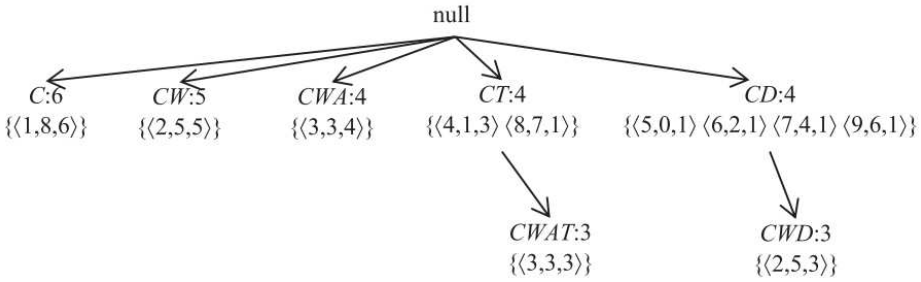


Figure 7. An example of the N-List algorithm for exploring repetitive transactions.

After executing the N-List structure, a set of repetitive transactions is extracted. Therefore, using the N-List structure, a set of value-added services is offered to new customers, based on the basket in the category designated for new customers.

Table 3. A sample of DB after deleting 1 single pattern and descending order.

Frequently ordered items	transaction
C, W, A, T	1
C, W, D	2
C, W, A, T	3
C, W, A, D	4
C, W, A, T, D	5
C, T, D	6

Results

Our focus is on customers of the Iranian telecommunications industry. Trials and simulations have been carried out on 10,000 telecommunication contacts. In the Table. 4, simulation performed in a system shown.

Table 4. Factors and Specifications of the Simulation System

Factor	properties
Disk size	500 GB
RAM memory	4 GB
The number of processors	Intel Core i5
Operating system	Windows 7

Evaluation criteria

This section generally reviews evaluation metric based on unsupervised and supervised algorithms. In the equations (7 to 10) the methods of calculating the accuracy, precision, recall and classification error are shown.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

In relation (4), TP (True Positive) denotes transactions that are class positive and classified as positive. TN (True Negative) denotes the number of transactions that are negative and classified as positive. FP (False Positive) also indicates the number of

transactions that were positive and classified as negative. Finally, FN (False Negative) shows transactions that are negative and classified as exactly negative. The equation to the validity and recall assessment is as follows.

$$precision = \frac{TP}{TP+FP} \tag{8}$$

$$ReCall = \frac{TP}{TP+FN} \tag{9}$$

Finally, the error rate is calculated by formula (10):

$$Error = 1 - (\frac{TP+TN}{TP+TN+FP+FN}) \tag{10}$$

Validation indices are used to measure the goodness of clustering results to compare between different clustering methods or to compare the results of a method with different parameters. Indicators for evaluating unsupervised learning techniques differ from supervised techniques. In this section, we introduce important indicators for credit evaluation based on internal and external validation indices.

Compactness, or CP, is the intrinsic data set of the dataset and is the first criterion for evaluating the goodness of the data separation based on the values and properties of the dataset. According to this criterion, data belonging to a cluster should be as close as possible to each other. The common criterion for determining data density is data variance. So a good clustering creates clusters of samples that are similar to each other. More precisely, this index calculates the average distance between each data pair according to the relation 9. X is a dataset consisting of a stream of x_i. Ω is a set of x_i

collected in a cluster. W is also a set of w_i that represents the center of Ω clusters. To measure the mean of the general index of compression in all clusters, we use the relation of 10 where k is the number of clusters obtained. Ideally, the members of each cluster should be as close as possible. Therefore, the lower the CP index, the better and higher the compression rate for clustering [19].

$$\overline{CP_i} = \frac{1}{|\Omega_i|}\sum_{x_i \epsilon \Omega_i} \|x_i - w_i\|. \|x_i - w_i\| = d(x_i . w_j) \tag{11}$$

$$\overline{CP} = \frac{1}{K}\sum_{i=1}^{k} \overline{CP_i} \tag{12}$$

(1). Separation Index (SP), which specifies the degree of separation between clusters. This index measures the Euclidean Distance between centers of the cluster using the equation (10), where SP is close to zero indicating closeness between the clusters.

$$\overline{SP} = \frac{2}{k^2 - k}\sum_{i=1}^{k}\sum_{j=i+1}^{k} \|w_i - w_j\|_2 \tag{13}$$

The Davies-Bouldin evaluation, or DB*:* introduced by Davis and Bouldin, two scientists in electricity in 1979, is not dependent on the number of clusters or the clustering algorithm. This criterion uses the similarity between two clusters (R_ji), which is defined by the dispersion of a cluster $\overline{CP_i}$ and the non-similarity between two clusters (d_ij). The similarity between the two clusters can be defined in different ways but must have the same equation conditions (14). The similarity of the two clusters is also measured using the relation (15) where the relations (16) measure d_ij.

$$
\begin{gathered}
R_i \geq 0 \\
R_{ij} = R_{ji} \\
if \overline{CP_i} = 0 \ and \overline{CP_j} = 0 \ then R_{ij} = 0 \\
if \overline{CP_j} > \overline{CP_k} and d_{ij} = d_{ik} then R_{ij} > R_{ik} \\
if \overline{CP_j} = \overline{CP_k} and d_{ij} < d_{ik} then R_{ij} > R_{ik}
\end{gathered} \tag{14}
$$

$$
R_{ij} = \frac{\overline{CP_i} + \overline{CP_J}}{d_{ij}} \tag{15}
$$

$$
d_{ij} = d(x_i . w_j) = \|x_i - w_i\| \tag{16}
$$

According to the material outlined and the similarity between the two clusters defined, the Davis Bouldin index is defined as a relation of (17), where R_i is calculated as a relation of (18). A DB value close to zero indicates that the clusters are compact and are spaced apart.

$$
DB = \frac{1}{k} \sum_{i=1}^{k} R_i \tag{17}
$$

$$
R_i = max_{j=1,\ldots,k \ and \ i \neq j} (R_{ij}) \tag{18}
$$

(2). The Dunn Validity Index (DVI) is similar to the cross-validation process used in supervised learning techniques (cross-validation is a model evaluation method that determines how generalizable the results of a statistical analysis on a data set are and Is independent of educational data.). It measures not only the degree of compression within the clusters but also the degree of dispersion between the clusters. Relation (19) defines this criterion.

$$DVI = \frac{min_{0<m\neq n<k}\left\{ min_{\substack{\forall\, x_i \in \Omega_m \\ \forall\, j \in \Omega_n}} \{\|x_i - x_j\|\} \right\}}{max_{0<m\leq k} max_{\forall\, x_i, x_j \in \Omega_m} \{\|x_i - x_j\|\}} \quad (19)$$

If the dataset contains separate clusters, the gap between the clusters is large (fraction) and its clusters (fraction denominator) are expected to be small. As a result, a larger value is more desirable for this criterion. The disadvantages of this criterion are time calculation and noise sensitivity (the diameter of the clusters can vary greatly if a noise data is available).

Dataset

In this article, we have used 10,000 contacts of Iran Telecom contacts database. This database has 14 attributes. Table. 5, shows the data features of the value added of Iranian telecommunication customers.

Table 5. Value Added Data Features of Iranian Telecommunication Customers.

Feature name	Description
msg_type	Transaction successful
mobile_no	Phone Number
txn_amount	Transaction amount
pr_code	Process Code
Rrn	
Response	Transaction successful
record_time	Transaction log time
bank_id	Bank ID
txn_type	Transaction type
target_mobile	Destination mobile number
topup_type	product type
Status	Transaction successful
Hour	Transaction log time
Capturedate	Transaction log date

Operating system used in this study Windows 7, operating system type also 32-bit operating system, 4GB RAM used - 3.06GB usable, Intel processor - Number of cores 7 (Core ™) i7 CPU) - Q 720 @ 1.60 GHz is 1.60 GHz.

Evaluation results

In this section, the results of accuracy, precision, recall and classification error of trusted customers for telecommunication company are analyzed by analyzing their basket using N-List algorithm without this algorithm and combining it with ensemble learning core. In this paper, we combine three deep neural networks algorithms, C4.5 decision tree and SVM-Lib support vector machine in order to analyze the portfolio and customer classification. Each of these algorithms has the properties shown in the tables (6-8). Table 4 shows the details of the deep neural networks algorithm for basket analysis and customer classification.

Table 6. Specifications for deep neural networks algorithm for basket analysis and customer classification.

Parameter	**Description**
Number of hidden layers	*53*
Core	*Core MML-ANN*
Entrance	*Cart analysis and customer classification*
Number of iterations of the algorithm	*10*
Output	*customer classification*
Number of threads	*1*
Type of neuron activation function	*Tanh Function (Hyperbolic Tangential Function (Scalable and Modified Sigmoid))*
Distribution function	*Gaussian function*
Number of training samples	*30%*
The type of network	*Artificial Neural Networks*

The table below shows the specifications of the C4.5 decision tree algorithm for analyzing the cart and customer classification.

Table 7. Characteristics of the C4.5 decision tree algorithm to analyze the portfolio and classify customers.

Parameter	Description
Entrance	*Training examples*
The core of the decision tree	*CoreGain-Ratio*
Output	*customer classification*
Maximum tree depth	*20*
Pruning the tree	*Able to prune the tree*
Confidence rate	*0.25*
Pruning the tree	*Predict tree pruning*
Minimum leaf size	*2*
Minimum size of tree leaf separation	*4*
Number of lessons per round	*3*

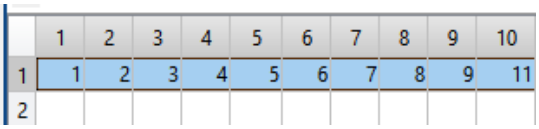
The following table shows the specifications of the SVM-Lib algorithm to analyze the portfolio and customer classification.

Table 8. Specifications of the SVM-Lib algorithm for portfolio analysis and customer classification.

Parameter	Description
Entrance	*Training examples*
Core support vector machine	*CoreLib*
Type of kernel	*RBF*
Output	*customer classification*
Backup vector type	*C-SVC(Two-class core type)*
Parameter C	*0.2*
Gamma	*0.3*
Epsilon	*0.001*

Therefore, the simulations are performed according to the features of each algorithm in accordance with the tables above. As described in Section 3, the N-List algorithm is used to select duplicate features. Repeatable features are those that are used by previous customers of Iran Broadcasting Company. After simulating the proposed method and implementing the N-List algorithm, the properties are selected as the iterative features shown in Fig. 8.

Figure 8. Final output of the N-LIST algorithm after applying it to the event logs.



	1	2	3	4	5	6	7	8	9	10
1	1	2	3	4	5	6	7	8	9	11
2										

As can be seen, the following properties have been extracted as N-LIST algorithms:

- msg_type attribute
- mobile_no feature
- txn_amount attribute
- pr_code feature
- Response feature
- record_time feature
- bank_id feature
- txn_type attribute
- target_mobile feature
- Status feature

In other words, the effective features of the basket analysis by the N-LIST algorithm are as follows.

Table 7. Effective Characteristics of Basket Analysis by the N-LIST Algorithm.

Row	Feature name	Description
1	msg_type	Message type
2	mobile_no	phone number
3	txn_amount	Transaction amount
4	pr_code	Process Code
5	Response	Response time
6	record_time	Transaction log time
7	bank_id	Bank ID
8	txn_type	Transaction type
9	target_mobile	Destination mobile number
10	Status	Transaction successful

Finally, the proposed hybrid algorithm is applied to the basket with these features and the results are discussed in the next section.

Analysis of Clustering Results

Before describing the results of the proposed method for basket analysis, this section examines the results of implementing clustering methods. Some of the most important metric to prove the validity of the K-Means clustering algorithm are:

(1). *CP*: The higher this criterion is, the more favorable the clustering will be.

(2). *SP:* The lower this criterion is, the better.

(3). *DB*: The higher this criterion is, the more favorable the clustering will be.

(4). *DVI*: The higher this criterion is, the more favorable the clustering will be.

These metric are discussed in the paper (Fahad et al, 2014). In order to prove the validity and desirability of the K-Means algorithm, the following section examines the mean of the metric derived from this algorithm with the other algorithms. Table. 8, shows a comparison of the compactness of the clustering methods with the K-Means algorithm.

The compression rate in Birch is 3.63, EM is 2.88, K-Means is 3.85, OptiGrid is 1.79 and Denclue is 1.35. K-Means-based algorithm outperforms other Birch, EM, OptiGrid and Denclue algorithms by 0.22, 0.97, 2.06 and 2.5, respectively. Fig.11, shows a comparison of the validity index of the clustering methods.

Table 8. Comparison of the compactness of the clustering methods with the K-Means algorithm

	CP	SP	DB	DVI
Birch	*3.63*	*0.57*	*5.78*	*4.11*
EM	*2.88*	*0.56*	*5.37*	*3.28*
OptiGrid	*1.79*	*0.52*	*4.27*	*2.16*
Denclue	*1.35*	*0.50*	*4.29*	***1.74***
XK-Means	***3.85***	***0.61***	***6.10***	*2.04*

The validation rate in Birch is 0.57, EM is 0.56, K-Means is 0.61, OptiGrid is 0.52 and Denclue is 0.501. Compared to other Birch, EM, OptiGrid and Denclue algorithms, the K-Means algorithm is 0.04, 0.05, 0.09 and 0.11, respectively. Fig. 12, shows the comparison of the Davis-Bouldin clustering methods. The DIV in Birch is 5.78, EM is 5.37, K-Means is 6.103, OptiGrid is 4.27 and Denclue is 4.29. Compared to other Birch, EM, OptiGrid and Denclue algorithms, the K-Means algorithm is 0.32, 0.73, 1.83, 1.81, respectively. Fig. 13, shows a comparison of the separation rates of the clustering methods. The separation index value in Birch is 4.11, EM is 3.28, K-Means is 2.04, OptiGrid is 2.16 and Denclue is 1.74. The K-Means algorithm improvement rate is 2.07 and 0.12, respectively, and worse than the Denclue algorithm compared to other Birch algorithms, EM algorithm, OptiGrid algorithm.

Basket Analysis Results Without the N-LIST Algorithm

This section analyzes the results of ensemble learning approaches such as deep neural networks, decision tree C4.5 and SVM-Lib in the form of an ensemble learning system. It should be noted that the N-LIST optimization algorithm can have a significant impact on improving the accuracy of the basket analysis. Therefore, in order to clarify and prove the effectiveness of the N-List algorithm in this section, we first analyze the results without this algorithm, then analyze the results obtained with this algorithm.

Calculating the accuracy, precision, recall and error analysis of the basket is one of the most important parameters that can prove the accuracy of the proposed method. Hence, Table. 9, shows the TP, TN, FP and FN results of the SVM-Lib, C4.5 and deep neural network algorithm for basket analysis with and without the N-LIST algorithm.

Table 8. TP, TN, FP and FN results of the SVM-Lib, C4.5 and deep neural network algorithm for basket analysis with and without the N-LIST algorithm.

	Without N-List algorithm				With N-List algorithm			
	SVM-Lib	C4.5	Deep	Ensemble	SVM-Lib	C4.5	Deep	Ensemble-Nlist
TP	7500	7650	7500	8200	7600	7560	7700	9240
TN	100	220	520	860	130	200	20	520
FP	200	90	2	90	120	140	80	50
FN	2200	2050	2000	850	2150	2100	2200	190

Table 8 shows the number of correct and incorrect classifications. Based on these variables, the criteria of accuracy, accuracy, recall, and error are calculated, which are described below. Table. 10, shows the accuracy, precision, recall and error rate of the SVM-Lib, C4.5 and deep neural network algorithm for basket analysis without the N-LIST algorithm.

Table 10. Accuracy, precision, recall and error rate of the SVM-Lib, C4.5 and deep neural network algorithm for basket analysis without the N-LIST algorithm

	Without N-List algorithm			
	Accuracy	Precision	Recall	Error
SVM-Lib	76	97.40	77.31	24
C4.5	78.62	98.83	78.86	21.37
Deep	80.02	99.97	78.94	19.97
Ensemble	**90.6**	**98.91**	**90.60**	**9.4**

As can be seen from the Table. 10, the accuracy of the ensemble learning algorithm without applying the N-List is 90.6%. The average improvement of classification accuracy and basket analysis in the proposed method is 12.38% compared to other algorithms. Also the precision, recall and error rate of the proposed method improved about 0.17%, 12.23% and 12.38% compared to other algorithms.

Table 11. Accuracy, precision, recall and error rate of the SVM-Lib, C4.5 and deep neural network algorithm for basket analysis with the N-LIST algorithm

	Without N-List algorithm			
	Accuracy	Precision	Recall	Error
SVM-Lib	80.3	99.08	80	19.7
C4.5	82.6	98.22	82.90	17.4
Deep	87.2	99.02	87.09	12.8
Ensemble	**97.6**	**99.44**	**97.94**	**2.4**

As can be seen from the Table. 11, the accuracy of the ensemble learning algorithm without applying the N-List is 97.6%. The average improvement of classification accuracy and basket analysis in the proposed method is 14.23% compared to other algorithms. Also the precision, recall and error rate of the proposed method improved about 0.66%, 14.61% and 14.23% compared to other algorithms.

Discussion

The value added service is of great benefit to telecommunications companies. Some customers also benefit from paying for VAS. In this paper, using supervised machine learning algorithms such as K-Means, ensemble learning algorithms consisting of a combination of deep neural networks algorithms, SVM-Lib, and C4.5 decision tree, as well as the N-List algorithm of basket analysis. Customers and classify customers who can make the most profit for telecommunications companies. By simulating the proposed method, it was observed that using N-List technique to extract the necessary features has a significant impact on the analysis of the customer's basket. Finally, by analyzing the customer cart with the proposed strategy, attractive services with acceptable accuracy were provided to customers.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets generated and/or analysed during the current study are not publicly available due [REASON WHY DATA ARE NOT PUBLIC] but are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests

Funding

Not applicable

Authors' contributions

All authors contributed to developing the ideas, and writing and reviewing this manuscript. All authors read and approved the fnal manuscript.

Acknowledgement

Not applicable

References

1. Jishya GB, Maran K. Influence of the Value Added Services (VAS) Consumer Decision with the Brand Names. Int. J Sup. Chain. Mgt. 2018 Feb;7(1):137.
2. Olya H, Altinay L, De Vita G. An exploratory study of value added services. J Serv Mark. 2018;
3. Chen MC, Chiu AL, Chang HH. Mining changes in customer behavior in retail marketing. Expert Systems with Applications. 2005 May 1;28(4):773-81.

4. Liu J, Gu Y, Kamijo S. Customer behavior classification using surveillance camera for marketing. Multimed Tools Appl. 2017;76(5):6595–622.

5. Kaur M, Kang S. Market Basket Analysis: Identify the changing trends of market data using association rule mining. Procedia Computer Science. 2016;85:78-85. https://doi.org/10.1016/j.procs.2016.05.180

6. Mansur A, Kuncoro T. Product inventory predictions at small medium enterprise using market basket analysis approach-neural networks. Procedia Econ Financ. 2012;4:312–20.

7. Haghighatnia S, Abdolvand N, Rajaee Harandi S. Evaluating discounts as a dimension of customer behavior analysis. J Mark Commun. 2018;24(4):321–36.

8. Kurniawan F, Umayah B, Hammad J, Nugroho SM, Hariadi M. Market Basket Analysis to identify customer behaviours by way of transaction data. Knowledge Engineering and Data Science. 2018;1(1):20.

9. Musalem A, Aburto L, Bosch M. Market basket analysis insights to support category management. Eur J Mark. 2018;

10. Szymkowiak M, Klimanek T, Józefowski T. Applying market basket analysis to official statistical data. Econometrics. 2018;22(1):39–57.

11. Valle MA, Ruz GA, Morrás R. Market basket analysis: Complementing association rules with minimum spanning trees. Expert Syst Appl. 2018;97:146–62.

12. Jain S, Sharma NK, Gupta S, Doohan N. Business Strategy Prediction System for Market Basket Analysis. in: Kapur P, Kumar U, Verma A. (eds) Quality, IT and Business Operations. Springer Proceedings in Business and Economics. Springer, Singapore; 2018. P 93-106.

13. Srivastava N, Stuti, Gupta K, Baliyan N. Improved Market Basket Analysis with

Utility Mining. InProceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT) 2018;26-27.

14. Jiang H, Kwong CK, Kremer GO, Park WY. Dynamic modelling of customer preferences for product design using DENFIS and opinion mining. Advanced Engineering Informatics. 2019 Oct 1;42:100969.

15. Venkatachari K, Chandrasekaran ID. Market Basket Analysis Using FP Growth and Apriori Algorithm: A Case Study of Mumbai Retail Store. BVIMSR's Journal of Management Research. 2016 Apr;8(1):56-63.

16. Sherly KK, Nedunchezhian R. A improved incremental and interactive frequent pattern mining techniques for market basket analysis and fraud detection in distributed and parallel systems. Indian J Sci Technol. 2015;8(18):1–12.

17. Ossama O, Mokhtar HMO, El-Sharkawi ME. An extended k-means technique for clustering moving objects. Egypt Informatics J. 2011;12(1):45–51.

18. Deng Z, Wang Z, Jiang J. A new algorithm for fast mining frequent itemsets using N-lists. Science China Information Sciences. 2012 Sep 1;55(9):2008-30.

19. Fahad A, Alshatri N, Tari Z, Alamri A, Khalil I, Zomaya AY, et al. IEEE TRANSACTIONS ON A Survey of Clustering Algorithms for Big Data : Taxonomy and Empirical Analysis. IEEE Trans Emerg Top Comput. 2014;2(3):267–79.

Figures

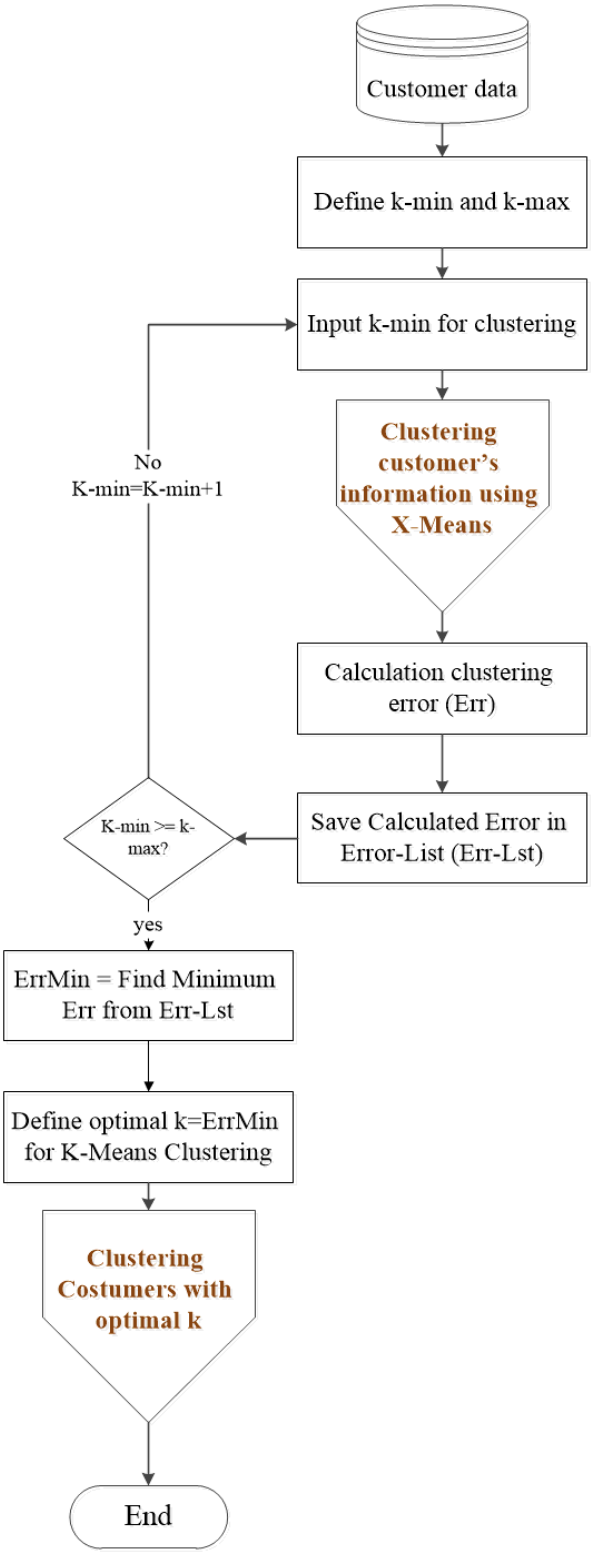


Figure 1

Flowchart of X-Means Clustering Algorithm for clustering customer's information

Algorithm: Extended k-Means (E-km)(S,C)

Input: S: List of segments in MOD, C initialized k cluster centroids, ¥ Give Threshhold
Output: C_{List}: List of Clusters
foreach (s ϵ S) **do**
foreach (c ϵ C) **do**
TempDist=Direction Evaluation (s, direction, c.
direction) + EuclideanDistance (s, c);
end
MinDistance=Min[TempDist]. centroid;
ClosetCentroid=Min[TempDist]. centroid;
if (MinDistance ≤ ¥) **then**
Cluster=C[ClosetCentroid];
C_{List}= Update.Centroid(Cluster, s);
else
$Cluster_{New}$. centroid=s;
C. Add ($Cluster_{New}$. centroid);
end
return$C_{List;}$

Figure 2

X-Means Clustering Algorithm [17].

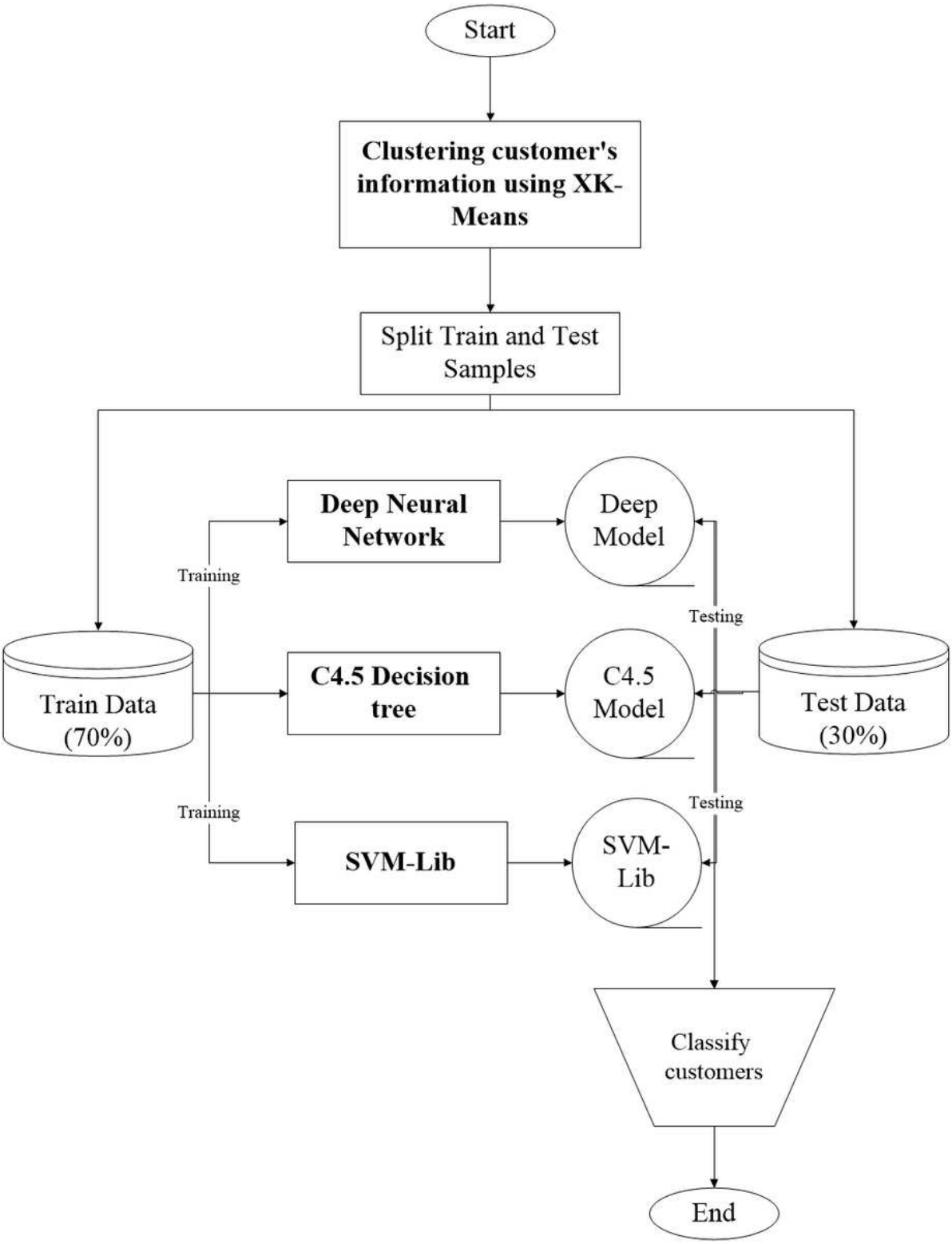


Figure 3

Ensemble learning flowchart for classify new customers

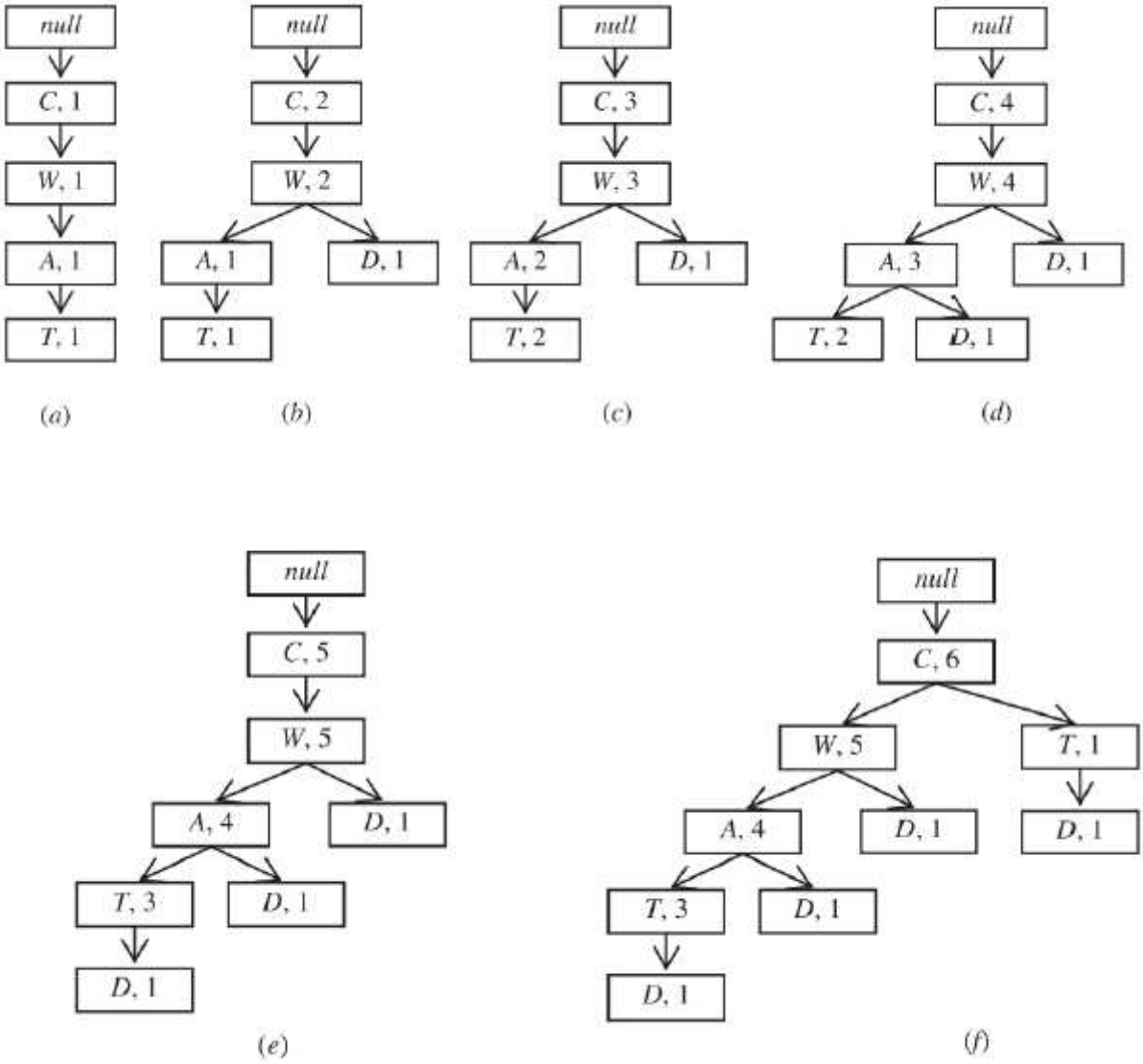


Figure 4

Illustration of PPC tree creation using DB example with minSup = 50%.

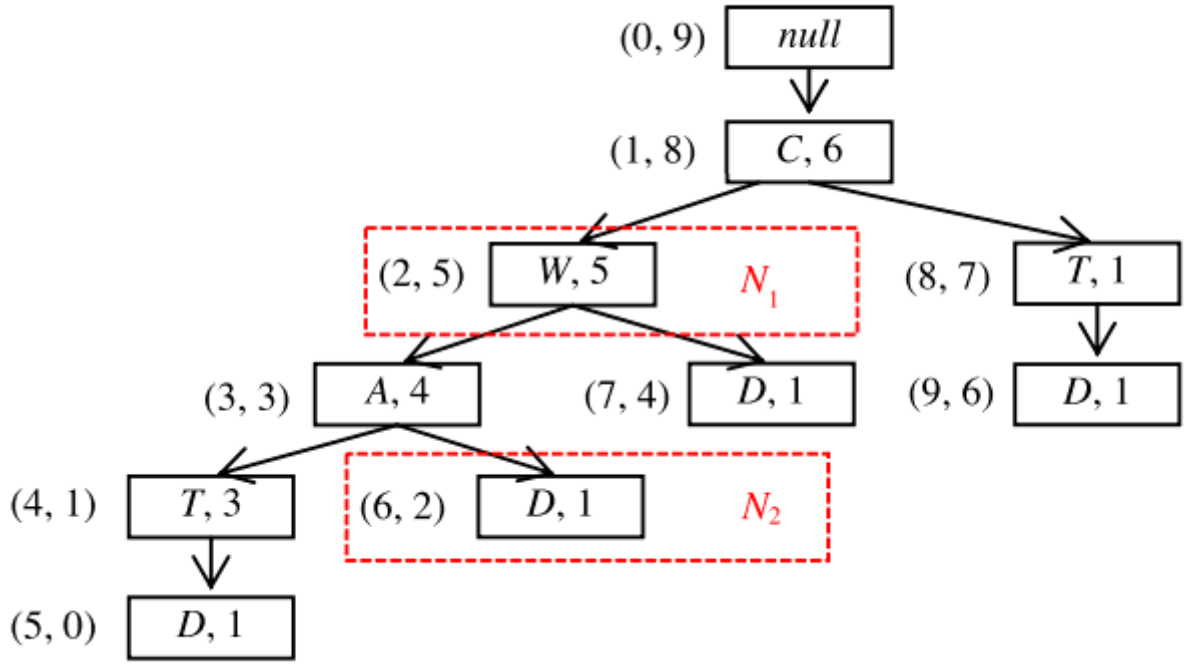


Figure 5

The final PPC tree created from the DB example with minSup = 50%.

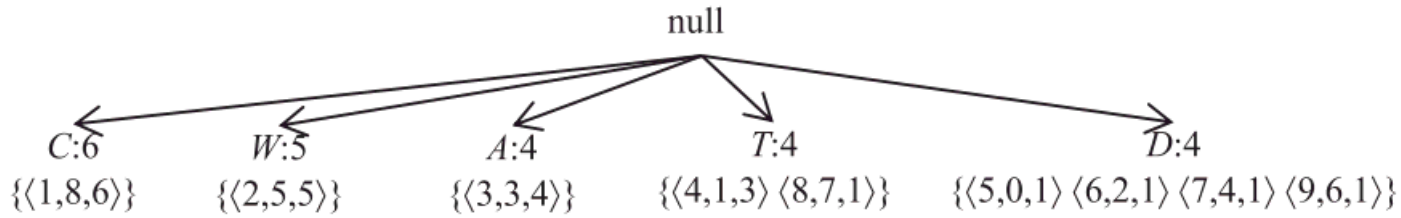


Figure 6

One-repetition frequency patterns and their N-List

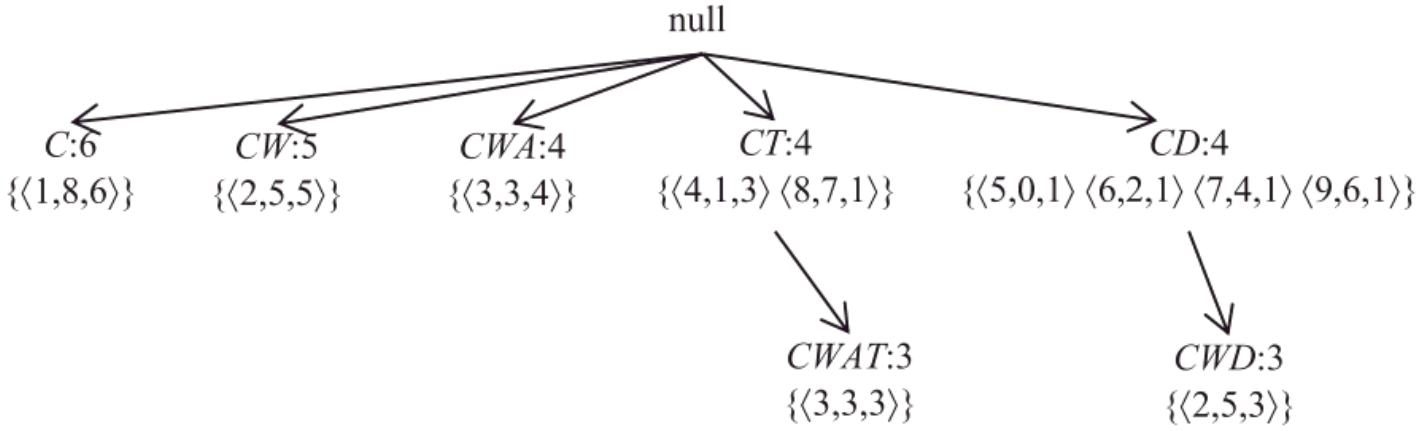
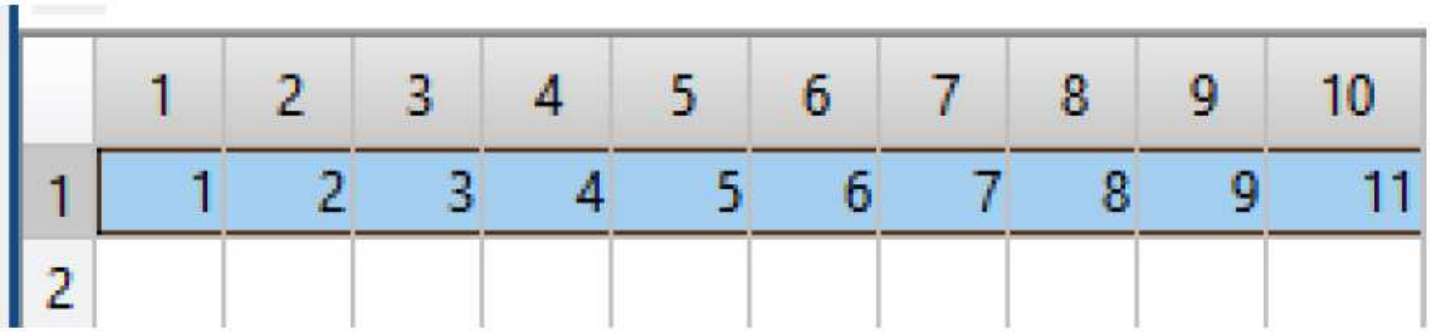


Figure 7

An example of the N-List algorithm for exploring repetitive transactions.



	1	2	3	4	5	6	7	8	9	10
1	1	2	3	4	5	6	7	8	9	11
2										

Figure 8

Final output of the N-LIST algorithm after applying it to the event logs