Journal of Big Data

RESEARCH Open Access

An analytics model for TelecoVAS customers' basket clustering using ensemble learning approach

Mohammadsadegh Vahidi Farashah[1], Akbar Etebarian[1*], Reza Azmi[2] and Reza Ebrahimzadeh Dastjerdi[1]

*Correspondence:
etebarian@khuisf.ac.ir
[1] Department of Management, Isfahan (Khorasgan) Branch, Islamic Azad University, Isfahan, Iran
Full list of author information is available at the end of the article

Abstract

Value-Added Services at a Mobile Telecommunication company provide customers with a variety of services. Value-added services generate significant revenue annually for telecommunication companies. Providing solutions that can provide customers of a telecommunication company with relevant and engaging services has become a major challenge in this field. Numerous methods have been proposed so far to analyze customer basket and provide related services. Although these methods have many applications, 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 telecommunication 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 telecommunication 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, the accuracy improved to about 7%.

Keywords: User behavior analysis, Basket analysis, Value added service, Ensemble learning, Deep learning

Introduction

Value-Added Services are one of the important features and capabilities of mobile telecommunication companies that enable customers to receive services by paying to mobile telecommunication companies. These services can be very useful and effective in analyzing customer behavior [1, 2]. Customer Behavior [3] in many online systems are the activities that a customer does on a regular 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 is top-selling. In addition, information can be used to add stocks to this sample. Also, transactions and customer performance can be used in the equation of every item present in the customer basket. Such information can also be used to present the right product to attract customers. One of the most important uses of these transactions is data analysis, transaction analysis and customer basket analysis [7].

Customer basket analysis is one of the modes of analysis based on customer behavior. However, shopping at the supermarket happens through identifying and making direct linkage among different items by 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 products that are repeatedly purchases, and sales and product alignment. 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 value-added service system of telecommunication companies.

There are still challenges and difficulties in providing services in value-added telecommunication systems, such as inadequate accuracy and high error of providing related services to the customers. Until now, there have been various methods for analyzing customers' portfolio, such as the method of customer basket analysis based on 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]. So, some of the most important motivations of this paper are as follows:

- Presenting a model for improving the analysis of Customer baskets and achieving the best use of the combination of algorithms in this case.
- Identifying the components and dimensions of the TelecoVAS Customer Basket model in Mobile Communication Company
- Extracting features of VAS Customer Basket
- Mobile communication companies and digital stores can use the obtained results to choose the appropriate solution to increase revenue, improve sales of services and products, and optimize their advertising and marketing.
- Universities and other scientific and research centers; Once the results of this research are known, they can use them to conduct more specialized research and rely on them in conducting further research in this field and developing scientific theories.

- Researchers and students; As the future makers of society, they should always have enough information about the analysis of the shopping cart of VAS subscribers in the mobile telecommunication company. Therefore, such research will help them to do other research.

In this paper, we used the N-List algorithm-based technique [14] 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. (Library core of the support vector machine algorithm) [15]. The proposed ensemble learning system is based on maximum votes and sends the best response to the output at each step.

The main contributions of this research can be summarized as follows:

- Combining K-Means and X-Means clustering algorithms and generating an efficient clustering algorithm to determine the initial grouping of data
- Combining ensemble learning method and N-List algorithm to analyze the TelecoVAS Customer Basket in Mobile Communication Company
- Combining machine learning methods in the ensemble learning system and improving the results obtained with the help of N-List algorithm

Therefore, by applying the N-list-based mining process technique in the model proposed in this paper and by maximizing it, more precise rules can be extracted by analyzing the TelecoVAS Customer Basket in the mobile telecommunication company. Therefore, providing an efficient N-list-based process-based modeling that can analyze customers' baskets is one of the most important aspects of this research innovation.

The remainder of this paper is presented as follows: "Related works" section reviews the work done in the past, "Proposed method" section describes the proposed approach and architecture. In "Results" and "Discussion and future suggestions" section the results are obtained and the final conclusions are discussed.

Related work

Some research about Customer basket analysis has been carried out. This section will refer to some papers that discussed the Customer basket analysis and outline models and technics.

In 2020, Seyedan et al. presented a classification of these algorithms and their applications [16].

Chain management to predict time series, clustering, K nearest neighbors, Neural networks, regression analysis, support vector machines and support vectors. This paper is based on meta-research demand forecasting in supply chains. Demand data features are expanding and dispersing today and Global supply chains use big data analysis and machine learning.

In 2020, Yudhistyra et al. proposed a method for implementing big data combining the CRISP-DM framework and key steps for customer analysis. This paper aimed to discover meaningful patterns and ensure high quality of knowledge discovery from the big data available in a company [17].

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 a dynamic model by using DENFIS approach. Unlike the conventional DENFIS approach that 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 [18].

In 2018, Musalem et al. 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 includes the supermarket level and has a poor performance in a 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 were 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 et al. 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 was 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 et al. used a portfolio optimization model of customer shopping in an efficient way called utility mining. They proposed utility mining as an improved model of data 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 for further development [13]. In 2017, Kurniawan et al. Presented a customer basket analysis model based on 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 was the speed of basket analysis. However, 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 did not perform well on large data sets [8]. In 2016, Kaur et al. 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 was sufficient accuracy in analyzing the customer basket. The proposed analysis was very slow and the model was 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 development [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 was based on sharing repeated transactions. One of the advantages of their approach was the improvement of accuracy and consistency of customer basket analysis. However, one of the major disadvantages of their method is the increased runtime and lack of potential for development in a larger statistical population [19]. In 2015, Sherly et al. used parallel and distributed techniques and associative rules to analyze the customer basket. In their research, they sought to increase the speed, completeness, and accuracy of customer basket analysis. Eventually they could enhance the accuracy to some extent and significantly improved the speed and comprehensibility through parallelization [20]. From the analysis of the research that has been done so far, it can be seen that most research suffers from inadequate accuracy, low speed of cart analysis, inadequacy and so on. Thus, considering the 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 network algorithms, C4.5 decision tree and SVM-Lib algorithm significantly enhances the quality of value-added services provided [15]. Reviewing the research that has been done so far, it can be seen that many of them suffer from inadequate accuracy, low speed of basket analysis, inadequacy, and so on. Thus, considering the 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 [14]. The process approach significantly enhances the quality of value-added services. Table 1 outlines the advantages and disadvantages of each method.

According to the review of the research conducted in the field of the TelecoVAS customer basket, it was found that the proposed methods face many challenges such as inaccuracy, recall, precision, and error rate. Therefore, in this paper we combined the X-Means algorithm, the ensemble learning system, and the N-List structure to analyze the customer portfolio of a mobile telecommunication 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 telecommunication company. The ensemble learning algorithm is also used to assign categories to new customers, and finally to the N-List structure for customer basket analysis.

The proposed method

The proposed method in this paper is based on X-Means clustering algorithms [21], N-List structure [14] for extracting frequent patterns, and ensemble learning system to provide attractive value added services to telecommunication customers. This section

Table 1 Comparison of the most important the advantages and disadvantages of the reviewed previous methods to an analytics model for customers' basket

Authors	Refs	Method	Advantages	Disadvantages
Seyedan and Mafakheri	[16]	Presented a classification of these algorithms and their applications	1. Good accuracy	1. Complex model
Yudhistyra and Risal	[17]	New method for implementing big data in this version combines the CRISP-DM framework and key steps for customer analysis.	1. High speed	1. Low accuracy 2. Low precision
Jiang et al.	[18]	New methodology for dynamic modelling of customer preferences based on online customer	1. Good accuracy	1. Low speed
Musalem et al.	[9]	The methodology also yields a segmentation of shopping trips based on the composition of each shopping basket	1. High speed	1. Low accuracy 2. Low precision
Szymkowiak et al.	[10]	An Apriori algorithm for customer basket analysis	1. Good accuracy	1. Complex model
Jain et al.	[12]	A statistical method to make the service closer to the person concerned	1. Normal accuracy	1. Complex model
Srivastava et al.	[13]	Used a portfolio optimization model of customer basket	1. High speed	1. Low accuracy 2. Low precision
Kurniawan et al.	[8]	Associative and data mining techniques such as neural networks and Apriori.	1. Fast Execution time	1. High MAE 2. High RMSE
Kaur and Kang	[5]	A customer basket analysis model using a combination of data mining methods and association rules	1. High speed	1. Low accuracy 2. Low precision
Venkatachari et al.	[19]	Used a combination of associative approaches such as Apriori and FP-Growth to analyze customer baskets	1. High speed	1. Low accuracy 2. Low precision
Sherly and Nedunchezhian	[20]	A parallel and distributed techniques and associative rules to analyze the customer basket	1. Normal accuracy	1. Low speed
Deng et al.	[14]	A process-based approach and algorithm for extracting iterative patterns such as N-List	1. Good accuracy	1. Complex model
Abdiansah and Wardoyo	[15]	Increases the accuracy of customer basket extraction and analysis. With the help of deep neural network algorithms, C4.5 decision tree and SVM-Lib algorithm	1. Good accuracy	1. Complex model

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

Data preprocessing

Once the data enters into the proposed system, inappropriate samples must be distinguished and removed as part of a preprocessing phase to keep data consistency. There are some popular ways to apply preprocessing to the data, such as:

- Clearing
- Collecting
- Transferring
- Reducing

In this paper, we use data clearing method. In the proposed strategy, we check the data, and if the row or column contains null or unused values, the mean of the next and previous values will be calculated and replaced with the null ones. Data clearing eliminates outliers and produces more consistent data.

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 telecommunication customers between [0,1]. In other words, all datasets are mapped into matrices, and matrix rows are normalized. Normalization is done due to achieve higher accuracy [22]. To normalize the values of each dataset, we use (1).

$$\text{Normalize}(x) = \frac{(\text{x} - \text{X}_{\text{min}})}{(\text{X}_{\text{max}} - \text{X}_{\text{min}})} \tag{1}$$

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]. Min–Max normalization is easier than other normalization methods and performs the normalization process faster. For this reason, the Min–Max normalization algorithm is used in this paper.

Customers clustering using XK-means algorithm

In this paper, we used a combination of K-Means [23] and X-Means algorithm together for clustering customer based on behavior information. The combination of K-Means and X-Means clustering algorithm is called XK-Means algorithm.

The second phase of this paper is of customers clustering. Customers may have two cases. The first case is a new customer who is active in the system. The other case is the one who is already registered in the system and has some activities. The X-Means clustering algorithm receives behavior information of customers. It then directs each customer to a cluster based on the behavior information. The X-Means algorithm is used to cluster the customer's information. One of the basic applications of using 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. Figure 1, illustrates the application of the X-Means clustering algorithm in clustering 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 of 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 [23] with the optimal K number is used for clustering.

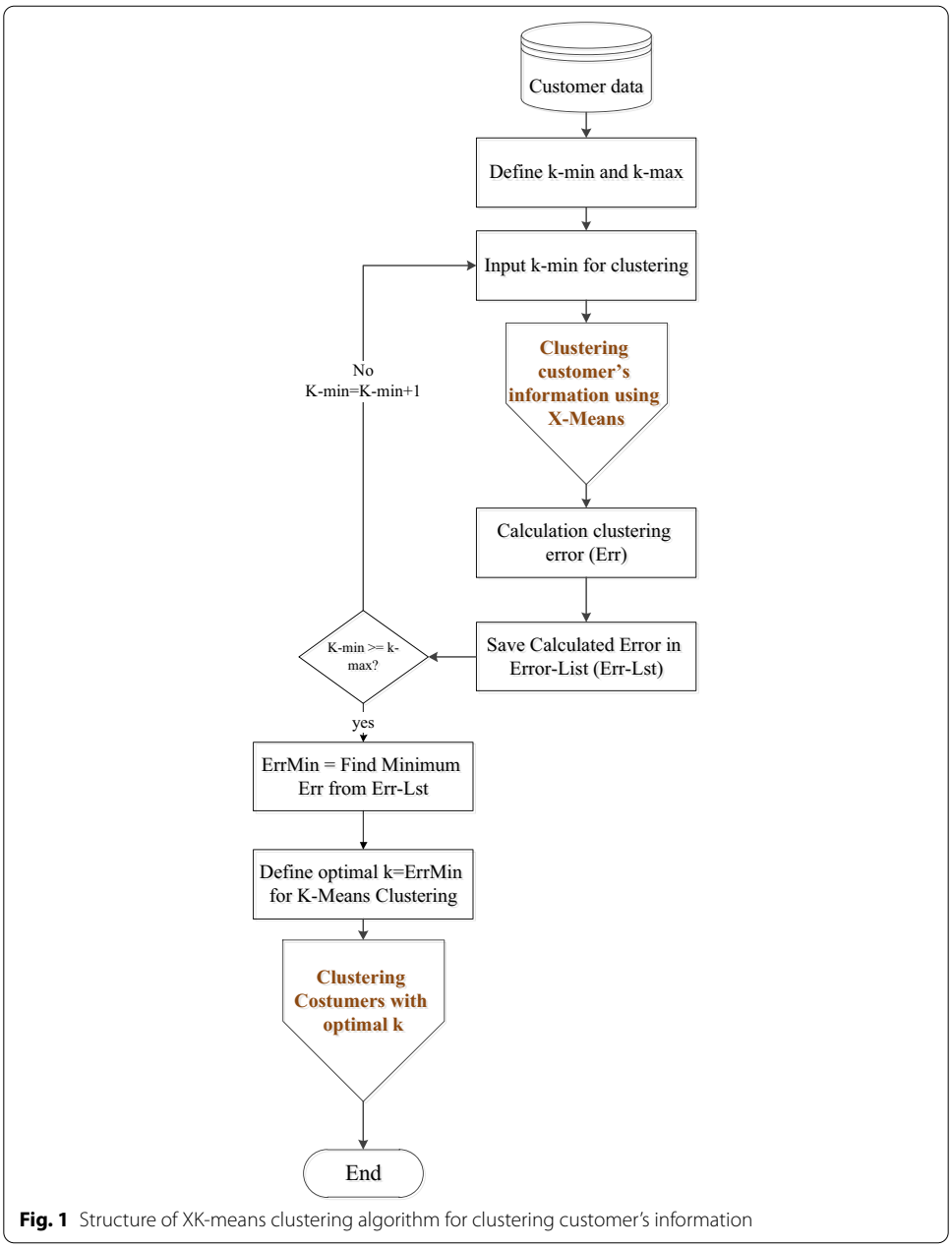

Fig. 1 Structure of XK-means clustering algorithm for clustering customer's information

The K-Means algorithm [23] 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 the clustering process is done based on the determined number. 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 number of *k* and determines the clusters with the lowest error rate as the basic clusters.

It uses this number of clusters as input to the K-Means algorithm and performs clustering of the customer's information. After the customer's information is clustered, the 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 (μ_1,...,μ_k).
3. Then repeat the following steps until the convergence becomes complete:

Calculate c for each i.

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

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

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

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.

As can be seen from Fig. 2, the initial k number is determined first. Then the K-Means clustering algorithm [23] 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-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 telecommunication 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 customers of the telecommunication company and

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$C_{List;}$

Fig. 2 X-means clustering algorithm [24]

labeling the customers is the output. Table 2 shows the sample of cluster and labeling customer's information.

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

The ensemble learning

Figure 3 shows the ensemble learning flowchart for classifying the customers. Figure 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 [15] for classifying new customers in mobile telecommunication 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 ultimate 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. Consider Test 1: a 35-year-old male customer who lives in X Province. This example is now entered into the Deep Learning Algorithm model and the category 1 is specified for the 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 [15] 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 20–30 and are male in X province. Thus the output of the ensemble learning system is as follows.

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

Table 2 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

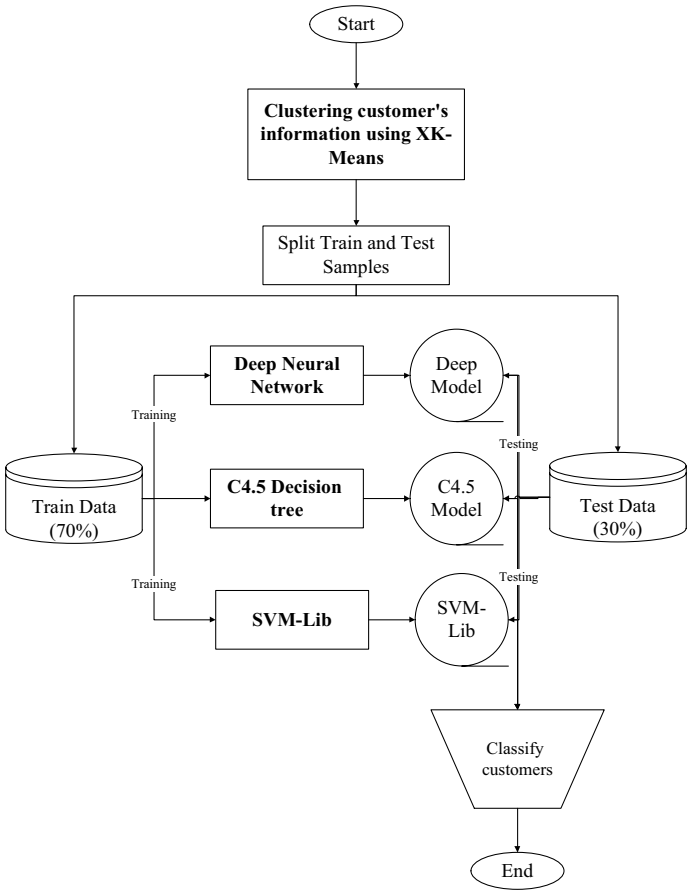

Fig. 3 Ensemble learning flowchart for classify new customers

structure [14] was implemented on all customer baskets in the selected category, and finally, based on the analysis, a set of services are provided for the new customers.

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 telecommunication system. In this study, the N-List associative algorithm is used to analyze the customer cart [14]. 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).

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 as ACW. Consider 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 given 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) [14]. 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:

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

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 4 illustrates the creation of a PPC tree [25] using the 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 employed 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 result in Table 4. In Fig. 5, the algorithm then, in turn, imports the remaining items in each transaction into the PPC tree.

Figures 6 and 7 shows 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 4).

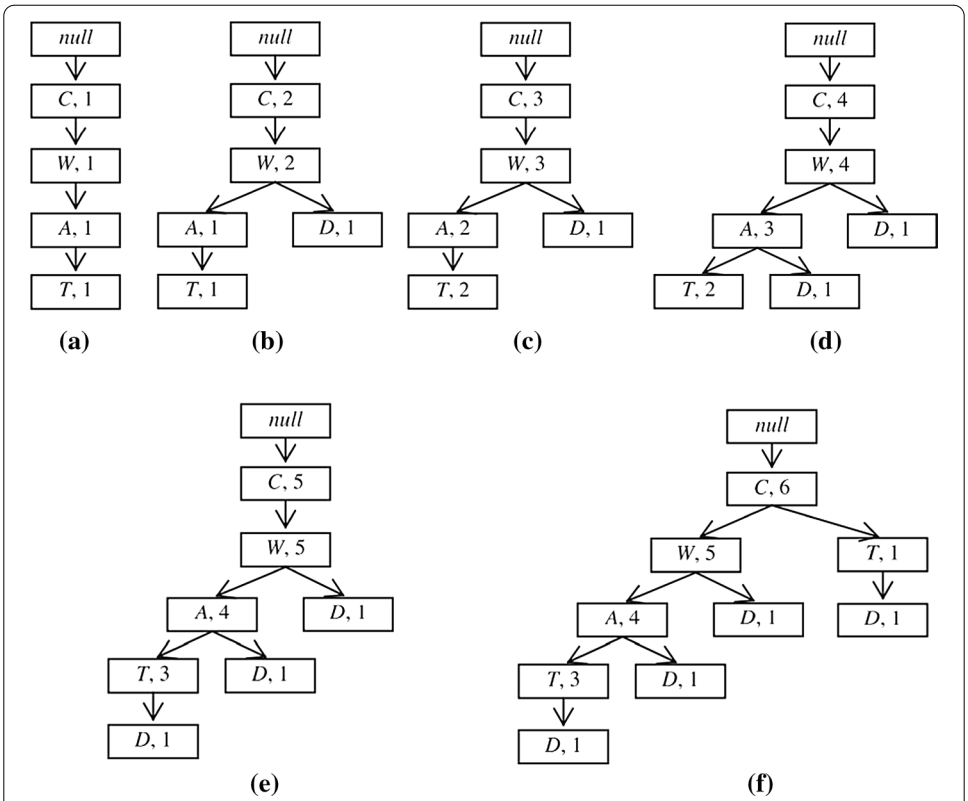

Fig. 4 Illustration of PPC tree creation using DB example with minSup = 50% [25]

Table 3 A sample DB dataset with six transactions [25]

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

Results

The proposed method is implemented using the MATLAB simulator version 2015a. The operating system is Windows 10 and of the 32-bit type. 4 GB of RAM is used from which − 3.06 GB is usable, with 7-core Intel processor (Core ™) i7 CPU)—Q 720 and processor base frequency of @ 1.60 GHz. Table 2 shows the settings and parameters of the network simulation. In this paper, we focused is on customers of the Iranian telecommunication industry. Trials and simulations have been carried out on 10,000 telecommunication contacts. In Table 5, simulation performed in a system shown.

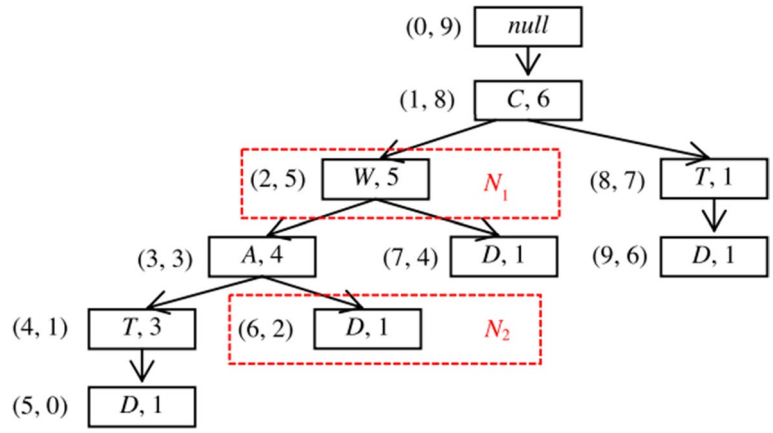

Fig. 5 The final PPC tree created from the DB example with minSup = 50% [25]

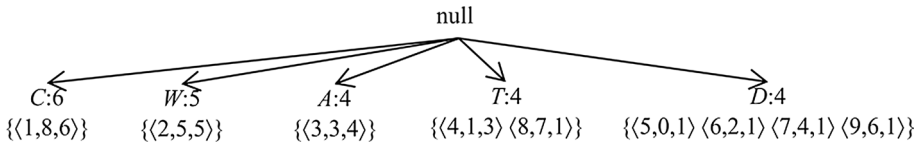

Fig. 6 One-repetition frequency patterns and their N-List [25]

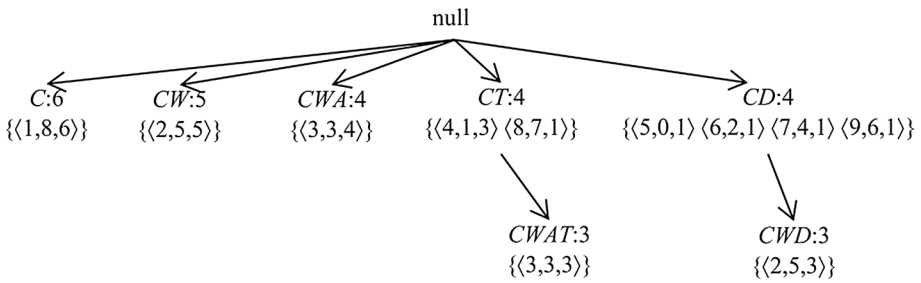

Fig. 7 An example of the N-List algorithm for exploring repetitive transactions [25]

Table 4 A sample of DB after deleting 1 single pattern and descending order [25]

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

Evaluation criteria

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

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

In Eq. (4), TP (True Positive) represents transactions that are positive and classified as positive.TN (True Negative) represents 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):

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

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

Compactness, or CP, related to the inherent information 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 the same cluster should be as close to each other as possible. 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

Table 5 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

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 [26].

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

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

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 Eq. (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.

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

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

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

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.

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

Table 6 Value added data features of iranian telecommunication customers

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

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

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 how it 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.

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

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 paper, we have used 10,000 contact information of Iran Telecom contacts database. This database has 14 attributes. Table 6, shows the data features of the value added of Iranian telecommunication customers.

The following section describes each of the dataset fields:

- msg_type: This property indicates the type of message.
- mobile_no: This feature shows the shared mobile number.

- txn_amount: This property shows the transaction amount.
- pr_code: This property shows the process code.
- rrn: This attribute indicates that the transaction was successful.
- response: This property indicates the response time.
- record_time: This feature shows the transaction registration time.
- bank_id: This property shows the bank ID.
- txn_type: This property indicates the type of transaction.
- target_mobile: This feature shows the service code.
- topup_type: This property indicates the type of product.
- status: This feature shows the status of the transaction.
- hour: This feature shows the transaction registration time.
- capturedate: This property shows the transaction registration date.

Operating system used in this study Windows 7, operating system type also 32-bit operating system, 4 GB RAM used—3.06 GB 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 7, 8, 9. Table 5 shows the details of the deep neural networks algorithm for basket analysis and customer classification.

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

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

Table 7 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	70%
The type of network	Artificial Neural Networks

Table 8 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

Table 9 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

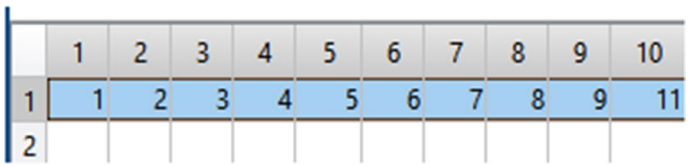

Fig. 8 Final output of the N-LIST algorithm after applying it to the event logs

Therefore, the simulations are performed according to the features of each algorithm in accordance with the tables above. As described in "Proposed method" section, 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.

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

- msg_type attribute
- mobile_no feature
- txn_amount attribute

Table 10 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

- 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 10).

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 the clustering will be.
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. [26]. 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. In this paper, we implemented the Birch [27], EM [28], OptiGrid [29] and Denclue clustering algorithm [30] and compared the result of proposed method (XK-Means) with these clustering algorithms.

Table 11 shows a comparison of the compactness of the clustering methods with the K-Means algorithm.

Table 11 Comparison of the compactness of the clustering methods with the K-Means algorithm

	CP	DVI	DB	SP
Birch [27]	3.63	0.57	5.78	4.11
EM [28]	2.88	0.56	5.37	3.28
OptiGrid [29]	1.79	0.52	4.27	2.16
Denclue [30]	1.35	0.50	4.29	*1.74*
XK-means (my approach)	*3.85*	*0.61*	*6.10*	2.04

Italic values indicate the best result

The compression rate in Birch is 3.63, EM is 2.88, XK-Means (The proposed method) 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. The second column shows a comparison of the validity index of the clustering methods.

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. The third column shows the comparison of the Davis-Bouldin clustering methods. The DB 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. The fourth column 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 improvement rate of K-Means based algorithm compared to Birch algorithms, EM algorithm, OptiGrid algorithm is 2.07, and 0.12, respectively, and it performed worse than Denclue 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 12, 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 12 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 13, 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.

As can be seen in Table 13, the accuracy of the ensemble learning algorithm without applying the N-List is 90.6%. The average improvement of classification accuracy and

Table 12 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 13 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

Table 14 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

	With 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

basket analysis in the proposed method is 12.38% compared to that of 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.

As can be seen from the Table 14, the accuracy of the ensemble learning algorithm with 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 that of 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 and future suggestions

The value-added services can make a great profit for telecommunication companies. Some customers pay for VAS to enjoy the services. In this paper, using unsupervised 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, customer's basket is analyzed and customers who

can have more profit for telecommunication companies are classified. By simulating the proposed method, it was observed that the use of N-List technique to extract repetitive features has a significant effect on customer basket analysis. So, one of the most important applications of the N-List algorithm in the proposed method is the extraction of repetitive rules. Extracting repetitive rules makes it possible to identify the top features in telecommunication services and use those features to analyze the subscriber portfolio.

The most important advantages of this article for other individuals and organizations are:

- Analyzing the TelecoVAS customer basket and providing attractive services for the targeted customers
- Increasing revenues from the TelecoVAS customer basket of the telecommunication company by finding and providing more accurate services related to customer needs than before.
- Increasing the number of users of the telecommunication company by offering services related to users' tastes.
- Decreasing costs of ads broadcasting and replacing it with the targeted ads based on the customer's taste.
- Customer grouping based on their tastes and needs. It is not only for TelecoVAS but also can be used for any dataset in digital markets.
- Extracting repetitive rules and processes in customers baskets and telecommunication company's focus on extraction rules.

Future research suggestions include the use of reinforcement learning methods and deep neural network LSTM, GMDH instead of mass learning system and also the use of other clustering algorithms such as DBScan instead of K-Means clustering algorithm in the clustering process.

Acknowledgement
Not applicable.

Authors' contributions
All authors contributed to developing the ideas, and writing and reviewing this manuscript. All authors read and approved the final manuscript.

Funding
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.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Department of Management, Isfahan (Khorasgan) Branch, Islamic Azad University, Isfahan, Iran. [2] Department of Engineering, Alzahra University, Tehran, Iran.

Received: 2 November 2020 Accepted: 29 January 2021
Published online: 17 February 2021

References

1. Gb J, Maran K. Influence of the Value Added Services (VAS) consumer decision with the brand names. Int J Sup Chain Mgt. 2018;7(1):137.
2. Olya H, Altinay L, De Vita G. An exploratory study of value added services. J Serv Mark. 2018;32:334–45.
3. Chen MC, Chiu AL, Chang HH. Mining changes in customer behavior in retail marketing. Expert Syst Appl. 2005;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 Comput Sci. 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 HS. 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. Knowl Eng Data Sci. 2018;1(1):20.
9. Musalem A, Aburto L, Bosch M. Market basket analysis insights to support category management. Eur J Mark. 2018. https://doi.org/10.1108/EJM-06-2017-0367.
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, editors. Quality, IT and business operations. Springer proceedings in business and economics. Singapore: Springer; 2018. p. 93–106.
13. Srivastava N, Stuti, Gupta K, Baliyan N. Improved market basket analysis with utility mining. In: Proceedings of 3rd international conference on internet of things and connected technologies (ICIoTCT); 2018. p. 26–7.
14. Deng Z, Wang Z, Jiang J. A new algorithm for fast mining frequent itemsets using N-lists. Sci China Inf Sci. 2012;55(9):2008–30.
15. Abdiansah A, Wardoyo R. Time complexity analysis of support vector machines (SVM) in LibSVM. Int J Comput Appl. 2015;128(3):28–34.
16. Seyedan M, Mafakheri F. Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities. J Big Data. 2020;7(1):1–22.
17. Yudhistyra WI, Risal EM, Raungratanaamporn IS, Ratanavaraha V. Using big data analytics for decision making: analyzing customer behavior using association rule mining in a gold, silver, and precious metal trading company in Indonesia. Int J Data Sci. 2020;1(2):57–71.
18. Jiang H, Kwong CK, Kremer GO, Park WY. Dynamic modelling of customer preferences for product design using DENFIS and opinion mining. Adv Eng Inform. 2019;42:100969.
19. Venkatachari K, Chandrasekaran ID. Market basket analysis using fp growth and apriori algorithm: a case study of mumbai retail store. BVIMSR's J Manag Res. 2016;8(1):56–63.
20. 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.
21. Pelleg D, Moore AW. X-means: Extending k-means with efficient estimation of the number of clusters, vol. 1. Inlcml; 2000. p. 727–34.
22. Kiran A, Vasumathi D. Data mining: min–max normalization based data perturbation technique for privacy preservation. In: Proceedings of the third international conference on computational intelligence and informatics. Singapore: Springer; 2020. p. 723–34.
23. Likas A, Vlassis N. The global k-means clustering algorithm. Pattern Recognit. 2003;36(2):451–61.
24. Ossama O, Mokhtar HMO, El-Sharkawi ME. An extended k-means technique for clustering moving objects. Egypt Inf J. 2011;12(1):45–51.
25. Le T, Vo B. An N-list-based algorithm for mining frequent closed patterns. Expert Syst Appl. 2015;42(19):6648–57.
26. 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.
27. Lorbeer B, Kosareva A, Deva B, Softić D, Ruppel P, Küpper A. Variations on the clustering algorithm BIRCH. Big Data Res. 2018;11:44–53.
28. Do CB, Batzoglou S. What is the expectation maximization algorithm? Nat Biotechnol. 2008;26(8):897–9.
29. Hinneburg A, Keim DA. Optimal grid-clustering: towards breaking the curse of dimensionality in high-dimensional clustering; 1999.
30. Rehioui H, Idrissi A, Abourezq M, Zegrari F. DENCLUE-IM: a new approach for big data clustering. Procedia Comput Sci. 2016;83:560–7.

Publisher's Note