Clustering on Ranked Data for Campaign Selection

UMANG GUPTA, VANDANA BHATTACHERJEE, AND PARTHA SARATHI BISHNU
Department of Computer Science and Engineering, Birla Institute of Technology, Ranchi 835215, India
Corresponding author: Umang Gupta (umanggupta7@gmail.com)

ABSTRACT
Recently the ranked data are commonly seen in the era of Internet and e-commerce where the consumers give their opinion in the form of ranks of a set of items online. The consumers are asked to put the ranks on items according to their order of preference. The applications of clustering ranked data are target marketing, campaign selection, top k-items, etc. The objective of this paper is to implement campaign selection process using clustering feedback data which are in rank ordered. To implement the proposed method we divide our experiments into two parts first, group the ranked data (consumer feedback), by applying different distance calculations, e.g., Kendall’s tau, Spearman’s rho square, Spearman’s footrule, Cayley’s distance. Second, use the knowledge derived from the groups in campaign selection process. We have compared our proposed clustering algorithm with existing algorithms on different real datasets, and results showed the effectiveness of our proposed algorithm.

INDEX TERMS
Clustering, marketing, opinion mining, ranked data.

I. INTRODUCTION
Target marketing is a popular tool by which companies are competing more effectively [1]. Using target marketing, companies are focusing to identify the group of those consumers they have the greatest chance of satisfying. Market segmentation, market targeting and market positioning are the basic components of target marketing. Dividing a market into well defined groups (where each group of consumers exhibits a similar set of tastes and desire) is called market segmentation. The task of the managers is to identify the right number of groups (segments) and the task of the datamining tool is to create the quality groups. Geographic, demographic, psychographic and behavioral are the segmentation variables through which one can divide the markets into segments. In behavioral segmentation, marketers split the consumers into groups on the basis of their products buying behavior, opinion and feedback towards the products, and knowledge and attitude toward the use of the products [1], [2]. Suppose we have a dataset which consists of the response of the consumers in ranked order. Using data mining tools, e.g., clustering algorithm, a manager can group those consumers who have similar taste and opinion. A campaign is a grouping of advertisement and each campaign can have multiple products for different market segments. Therefore, companies create separate campaigns for different market segments with different advertising objectives [1].

Now a days recording rank order questions on items are one of the most popularly used survey question types. The advantages of recording rank order scale are higher response rate and ease of use since the questions are close ended and consumers can give their responses quickly and effectively without wasting much time. But just like other survey questions, rank order questionnaire too have limitations e.g., missing ranks or partially filled the questionnaire. In such scenarios we may take up actions like filling missing ranks with (a) average of the rest of the ranks, (b) randomly assigned rest of the ranks, (c) delete the data with missing ranks, (d) use statistical or soft computing tools to guess the missing ranked values [3]–[6]. In this paper, we consider data without missing ranks. The shopkeepers, retailers, data brokers record consumer data physically which help producers, service providers, marketers to understand the requirements and preferences of a single consumer or groups of consumers. Companies record personal data physically where as ranking items is one of the common technique to create consumer data. Online e-commerce companies produce and use product’s ranking given by consumers [7].

A dataset \( F \) with ranked data is defined by a set of consumer feedbacks \( \{ f_1, f_2, \ldots, f_n \} \) where \( n \) is the number of
consists of rank values \( \{\pi_{1}, \pi_{2}, \ldots, \pi_{d}\}\), of the items \( I_{j}\), \( 1 \leq i \leq n, 1 \leq j \leq d, I_{j} \in \Theta\). The ranking \( \pi_{ij}\) is a permutation on the set of available items \( d \) i.e., \( \pi_{ij} \in (1, 2, \ldots, d)\). Consumers give their feedback on \( d \) number of products. For example, let the feedback of 2\textsuperscript{nd} consumer be \( f_{2} = \{1, 2, 4, 3\}\) (Table 2) then the rank of the 1st product \( f_{21}\) is \( 1 (\pi_{21})\), 2nd product \( f_{22}\) is \( 3 (\pi_{22})\), 3rd product \( f_{23}\) is \( 4 (\pi_{23})\) and 4th product \( f_{24}\) is \( 3 (\pi_{24})\). For \( f_{1}, f_{j} \in F\) where \( f_{1} = \{I_{1}, I_{2}, \ldots, I_{d}\}\) and \( f_{j} = \{I_{j}, I_{j}, \ldots, I_{j}\}, 1 \leq i, j \leq n\). If \( \pi_{ik} > \pi_{jk}\) then \( k\)\textsuperscript{th} consumer gives more preference on \( k\)\textsuperscript{th} item than \( j\)\textsuperscript{th} consumer gives. In Table 2 for feedback \( f_{1}\), \( f_{j} = \{I_{j}, I_{j}, \ldots, I_{j}\}\) i.e., consumer 1 gives the same feedback data on all products. The distance between \( f_{1}\) and \( f_{j}\) is defined as \( dist_{M}(f_{1}, f_{j})\). Since the data consists of ranked values hence we use four special types of distance calculations as follows: (a) Spearman’s footrule (footDist) [8]–[10] (b) Sperman’s rho square distance (smDist) [9], [11] (c) Kendall’s tau (kenDist) [9], [10], [12] and (d) Cayley’s (cayleyDist) [9], [13].

### TABLE 1. Consumer feedbacks.

| \( f_{j} \) | \( l_{1} \) | \( l_{2} \) | \( l_{3} \) | \( l_{4} \) |
|-----------|-----------|-----------|-----------|-----------|
| \( f_{1} \) | \( l_{11}(\pi_{11}) \) | \( l_{12}(\pi_{12}) \) | \( l_{13}(\pi_{13}) \) | \( l_{14}(\pi_{14}) \) |
| \( f_{2} \) | \( l_{21}(\pi_{21}) \) | \( l_{22}(\pi_{22}) \) | \( l_{23}(\pi_{23}) \) | \( l_{24}(\pi_{24}) \) |
| \( f_{3} \) | \( l_{31}(\pi_{31}) \) | \( l_{32}(\pi_{32}) \) | \( l_{33}(\pi_{33}) \) | \( l_{34}(\pi_{34}) \) |
| \( f_{n} \) | \( l_{n1}(\pi_{n1}) \) | \( l_{n2}(\pi_{n2}) \) | \( l_{n3}(\pi_{n3}) \) | \( l_{n4}(\pi_{n4}) \) |

In this research, we have considered these four distance measures for clustering ranked data. Comparative study of all the distance measures based on evaluation parameter are presented in section V. Now we describe various distance calculations as follows:

### A. SPEARMAN’S FOOTRULE DISTANCE [8], [9]

Let the data \( f_{1} = \{I_{1}, I_{2}, \ldots, I_{d}\}, f_{j} = \{I_{j}, I_{j}, \ldots, I_{j}\}\) and the content of the \( f_{1} = \{\pi_{1}, \pi_{2}, \ldots, \pi_{d}\}\) and \( f_{j} = \{\pi_{j}, \pi_{j}, \ldots, \pi_{j}\}\) are the ranks. First of all arrange the ranks of the attributes of \( f_{1}\) in ascending order of the rank values and then re-arrange the \( f_{j}\) ranks accordingly and it is applicable for all the four techniques. The \( dist_{footDist}(f_{i}, f_{j}) = \frac{\psi_{ij}}{max_{dist}}\), where \( \psi_{ij} = \sum_{p=1}^{d} |(\pi_{ip} - \pi_{jp})|\). If the dimension \( d \) is an odd number then \( max_{dist} = 2 \sigma (\sigma + 1)\), where \( \sigma = (d - 1)/2\). If the dimension \( d \) is an even number then \( max_{dist} = 2 \sigma^{2}\), where \( \sigma = d/2\).

### B. SPEARMAN’S RHO SQUARE DISTANCE [9], [11]

The \( dist_{smDist}(f_{i}, f_{j}) = \sqrt{\frac{\sum_{p=1}^{d} (\pi_{ip} - \pi_{jp})^{2}}{max_{dist}}}, \) where \( max_{dist} = \sqrt{\sum_{p=1}^{d} \sum_{q=1}^{d} (\pi_{pq} - \pi_{jq})^{2}}\), and \( 0 \leq dist_{smDist}(f_{i}, f_{j}) \leq 1\). When the ranks of both the feedback data \( f_{i}\) and \( f_{j}\) are in the same order then \( dist_{smDist}(f_{i}, f_{j}) = 0\) and when the ranks of both the data \( f_{i}\) and \( f_{j}\) are in reverse order then \( dist_{smDist}(f_{i}, f_{j}) = 1\).

### C. KENDALL’S TAU DISTANCE [9], [12]

The \( dist_{kenDist}(f_{i}, f_{j}) = \frac{T_{ij}}{max_{dist}}\), where \( max_{dist} = d(d - 1)/2\), and \( T_{ij} = \sum_{p=1}^{d} \sum_{q=p+1}^{d} k_{pq}^{ij}\), where \( k_{pq}^{ij} = 1\), if \( \pi_{ip} > \pi_{jq}\) and \( k_{pq}^{ij} = 0\), if \( \pi_{ip} \leq \pi_{jq}\). The \( dist_{kenDist}(f_{i}, f_{j}) = 0\) if \( f_{i}\) and \( f_{j}\) are in the same order (or the ranks are identical) and \( dist_{kenDist}(f_{i}, f_{j}) = 1\) if the ranks of \( f_{i}\) and \( f_{j}\) are in the reverse order.

### D. CAYLEY DISTANCE [9], [13]

The calculation is \( dist_{cayleyDist}(f_{i}, f_{j}) = \sum_{l=1}^{d-1} c_{l}\), where \( c_{l}\) is the swapping value. The swapping value of the \( i\)\textsuperscript{th} attribute is \( c_{l} = 1\) if \( \pi_{il} > \pi_{lj}\) and swap the values in between \( i\)\textsuperscript{th} and \( j\)\textsuperscript{th} index (or item) of the feedback. The \( c_{l} = 0\) if \( \pi_{il} \leq \pi_{lj}\), where \( \pi_{l}\) is the rank value of the \( i\)\textsuperscript{th} attribute and \( \pi_{l}\) is the rank of the \( p\)\textsuperscript{th} attribute of the same feedback data and do not swap the values in between \( i\)\textsuperscript{th} and \( j\)\textsuperscript{th} index.

The main objectives of this paper are to devise a campaign selection process using clustering algorithm and the steps which we have carried out as follows: first, (a) apply different distance measures to cluster the ranked data (consumer data), (b) propose a new initialization method for selection of initial cluster centers and (c) apply a greedy method to identify the cluster centers if the content (rank) of the cluster center is repeated; second, in the campaign selection process, using generated clusters of consumer data, (d) we can identify the item sets for the given target groups and (e) we can identify the target group for the given campaign. We have proposed three algorithms, KR1, KR2 and KR3. In KR1 algorithm, the initial cluster centers are selected randomly. In KR2, the initial cluster centers are selected based on our proposed technique, since to the best of our knowledge, no other researcher has proposed the initialization of clustering centers with rank data. In both KR1 and KR2, simple mode is calculated to update the cluster centers. But in these two methods the cluster centers may consist of repeated ranks. Therefore, we propose our third algorithm KR3. In KR3, the initial clusters centers are selected based on our proposed technique and if any center contains repeated rank then to make content of the center distinct our proposed non-repeated
cluster center identification technique is applied. We have compared our proposed techniques KR1, KR2 and KR3 with existing techniques K-Means (KM), K-Mediod (KMe) and K-Mode (KMo) on synthetic and real datasets and the results are presented.

The outline of this paper is as follows: in Section 2 we present related work, in Section 3 we describe clustering algorithms for consumer ranked data and campaign selection process. In Section 4 we explain the experiments. In Section 5 we present the results and analysis and this is followed by conclusions in Section 6.

II. RELATED WORK
In this section, we discuss the related work for both clustering algorithm and campaign selection process in marketing. In the field of clustering, Werrij and Kaptein [14], suggested K-Means, spectral clustering, and evolutionary algorithm for grouping ranked order data. But since the data are ranked ordered, hence it is not appropriate to use the mean value as cluster centers. Brentari et al. [2], used a weighted Spearman’s rho method for dissimilarity measure. Jacques and Biernack [5], suggested model-based clustering algorithm on multivariate partial ranking data. This is an extension of the “insertion sorting rank model” for ranking data, which has the dual property to be a meaningful model through its location and scale parameters. Grbovic et al. [15], proposed supervised clustering in the context of label ranking data using various baseline algorithms. Taritsiano [16], discussed hierarchical based clustering algorithm, where the author suggested weighted correlation coefficient for distance calculations. Busse et al. [6], suggested clustering of heterogeneous rank data using EM algorithm. Viappiani [17], suggested rank aggregation to cluster rank data. Most recently Peng et al. [22], suggested deep clustering method by minimizing the discrepancy between sample assignments with respect to multiple distance metrics. In [23] authors proposed a method to cluster multi-view data without parameter selection. Peng et al. [24], presented a novel deep learning model structured AutoEncoders for subspace clustering.

Now we discuss about related works on marketing management using data mining. Chan [18], discussed a novel approach that combines consumer targeting and consumer segmentation for campaign strategies. The author proposed a generic algorithm select appropriate consumers for each campaign strategy. Paetz [19], proposed a process for product rankings for each consumer using rule-based classification, genetic algorithm and k-NN algorithm. The basic differences between this work and our proposed work are firstly, the authors have proposed a technique using different classification techniques for the identification of ranks of the product for each consumer, but in our case consumers give their preferences on items and from the rank data we identify the clusters. Secondly, the authors have used modified Canberra distance for clustering which is time inefficient, to overcome these drawbacks we may use Kendall’s tau, Spearman’s rho square, Spearman’s rho, Spearman’s footrule, Cayley’s distance which are appropriate for distance calculation of ranked data. Fleming and Pashekevich [20], proposed a multi-objective algorithm approach for the problem of optimal TV advertising campaign generation for multiple brands using genetic algorithm. Michel et al. [21], proposed an effective selection of consumers for direct marketing campaigns based on net scores.

III. CLUSTERING ALGORITHM FOR CONSUMER RANKED DATA
First, we present the symbol table and then describe our proposed clustering techniques:

\[ P = \text{Consumer feedback dataset, } f_i \in P \]
\[ \theta = \text{Set of items for campaign selection} \]
\[ f_i = \text{The } i^{th} \text{ consumer feedback} \]
\[ I_{ij} = \text{The } j^{th} \text{ attribute of } i^{th} \text{ feedback} \]
\[ \pi_{ij} = \text{The rank of } j^{th} \text{ attribute of } i^{th} \text{ feedback} \]
\[ d = \text{Number of items in each feedback} \]
\[ \text{dist}_M(f_i, f_j) = \text{The distance between } f_i \text{ and } f_j \]
\[ n = \text{Number of feedbacks} \]
\[ D_T = \text{Set of n distances} \]
\[ D_B = \text{Set of distinct distances, } D_B \subseteq D_T \]
\[ k = \text{Number of clusters} \]
\[ k_i = \text{Number of items for the forth coming campaign} \]
\[ C_i = \text{ } i^{th} \text{ cluster} \]
\[ c_i = \text{Cluster center of } C_i \text{ cluster} \]
\[ q_i = \text{Number of data present in } C_i \text{ cluster} \]
\[ CG = \text{Set of campaigns } \{cq_i \} \subseteq CG \]
\[ s = \text{Total number of campaigns i.e., } |CG| \]
\[ n_i = \text{Number of items present in the } i^{th} \text{ campaign} \]
\[ LAF = \text{Local attention factor} \]
\[ FAF = \text{Final attention factor} \]
\[ GAF = \text{Grand attention factor} \]

To form the consumer group we calculate the similarity between the data. Here, we see our first definition “more similar data” as follows:

**Definition 1:** More similar data : Let three feedback be \( f_i, f_j \) and \( f_k \) and \( \text{sim}_M(f_i, f_j) \geq \text{sim}_M(f_i, f_k) \) or \( (1 − \text{dist}_M(f_i, f_j)) \geq (1 − \text{dist}_M(f_i, f_k)) \), \( M \in \{\text{Spearman’s footrule, superman rho square distance, Kendall’s tau distance, Cayley’s distance}\} \) then \( f_j \) is more similar data of \( f_i \) and vice versa.

**A. IDENTIFICATION OF THE INITIAL CLUSTER CENTERS**

The quality of the partitional clustering algorithms (e.g., K-Means algorithm) are dependent on the selection of the initial cluster centers. Randomly selecting the initial cluster center is the simple and widely used method. But there is no guarantee that one will not choose two seeds near the center of the same cluster and repeated run with randomly selected initial cluster selection increases overall time complexity [25], [26]. Here we propose a method to identify the initial cluster center for ranked data to ensure that chosen initial cluster centers are apart enough, we use the steps as follows:

**Step 1:** Calculate distance \( \text{dist}_M(f_i, f_j) \in D_T, 1 \leq j \leq n \), \( n \) is the number of data, \( f_j \in F \), where the rank of the
attributes of $f_x$ (or $f_y$) data (selected as pivot data) are in ascending (or descending) order, i.e., $f_x = \{1, 2, \ldots, d\}$ (or $f_y = \{d, d-1, \ldots, 1\}$).

**Step 2**: $D_D$ contains the distinct distances, where $D_D \subseteq D_T$ and arrange $D_D$ in ascending order of distances.

**Step 3**: From the $n$ number of distances we decide to select the initial $k$ cluster centers as follows: If $k = 2$, select the two data having minimum and maximum distances in $D_T$ as initial cluster centers. If $k = 3$, select the two data having minimum and maximum distances and a single data having median distance (from $D_T$) as initial cluster centers. Note that we can take lower or upper median value for even number of distances. If $k > 3$, select the two data having minimum and maximum distances and rest of the $t_i$ data from the $D_D$, where $t_i = \lceil \frac{D_D[x]}{k} \rceil$, $1 \leq i \leq k - 2$, and $k$ is the number of clusters, as initial cluster centers.

**Explanation**: All the data in the dataset must be within the rank format ($f_x = \{1, 2, \ldots, d\}$ to $f_y = \{d, d-1, \ldots, 1\}$ and the distance between $\{1, 2, \ldots, d\}$ and $\{d, d-1, \ldots, 1\}$ is the maximum. Hence, using $f_x$ (data with ascending ordered ranks) or $f_y$ (data with descending ordered ranks), we calculate the distances of the $f_j$, $1 \leq j \leq n$, to identify the range of the distances. If $k = 2$, we select the two extreme data as initial cluster centers [27]. Similarly, for the initial cluster centers $k > 2$, we select the data as initial data in such a way, that the initial cluster centers are far apart.

**Example 1**: Selection of initial cluster centers

Let there be seven feedback $F = \{f_1, f_2, f_3, f_4, f_5, f_6, f_7\}$ of the consumers as follows $\{(1, 2, 3, 4); (1, 2, 3, 4); (1, 2, 3, 4); (2, 3, 1, 4); (2, 3, 1, 4); (3, 1, 2, 4); (4, 3, 2, 1)\}$ respectively. The distances from $f_x = \{1, 2, 3, 4\}$ using Kendall’s tau distance (to explain the example, we have selected Kendall’s tau distance) are $D_T = [0, 0, 0.1667, 0.3333, 0.3333, 0.5, 1]$ respectively and distinct distances are $D_D = [0, 0.1667, 0.3333, 0.3333, 0.5, 1]$. If $k = 2$, select the two data having minimum (here, 0) and maximum (here, 1) distances, i.e., select $1^{st}$ data $(1, 2, 3, 4)$ and $7^{th}$ data $(4, 3, 2, 1)$ as initial cluster centers.

If $k = 3$, select the two data having minimum (here, 0) and maximum (here, 1) distances and the data having median distance (here, 0.3333) distance, i.e., select data, $1^{st}$ data $(1, 2, 3, 4)$, $7^{th}$ data $(4, 3, 2, 1)$ and $4^{th}$ or $5^{th}$ data $(2, 3, 1, 4)$ as initial cluster centers.

Similarly, if $k = 4$, select the two data having minimum and maximum distances and rest of the two data $t_1 = \lceil \frac{D_D[x]}{4} \rceil = \lceil \frac{2.5}{4} \rceil = 1.25 = 2^{nd}$ i.e., $2^{nd}$ data from the $D_D$ i.e., and $t_2 = \lceil \frac{D_D[x]}{4} \rceil = \lceil \frac{2.5}{4} \rceil = 2.5$, i.e., $3^{rd}$ data from the $D_D$, i.e., select the data, $1^{st}$ data $(1, 2, 3, 4)$, $7^{th}$ data $(4, 3, 2, 1)$ (from $D_D$), $2^{nd}$ data $(1, 3, 2, 4)$, and $3^{rd}$ data $(2, 3, 1, 4)$ (from the $D_D$) as initial cluster centers. Note that last two data are from $D_T$ (i.e., $3^{rd}$ and $4^{th}$ data of $D_T$).

**B. IDENTIFICATION OF CLUSTER BELONGNESS, AND CLUSTER CENTERS**

Here we discuss about the cluster belongingness and cluster center identification as follows:

**Definition 2**: Cluster belongingness: Data $f_i$ should belong to $r^{th}$ cluster, if and only if $\text{sim}(f_i, c_j) \geq \text{sim}(f_i, c_s)$, $1 \leq r, s \leq k$, $r \neq s, c_r$ and $c_s$ are the clusters centers of the clusters $C_r$ and $C_s$ respectively.

**Definition 3**: Cluster centers: A cluster center ($c_j$) of cluster $C_j = \{f_1, f_2, \ldots, f_q\}$, where $q$ is the number of data present in the $j^{th}$ cluster, is a vector $c_j = \{I_{f_1}, I_{f_2}, \ldots, I_{f_q}\}$ where $\sum_{i=1}^{q} (\text{dist}_M(f_i, c_j))/q$ is the minimum.

**C. THE KRanks CLUSTERING ALGORITHM**

Now we formally discuss about the simple KRanks clustering algorithm (similar to $k$-Means clustering algorithm) as follows:

**Algorithm 1 KRanks Clustering Algorithm**

**Input**: Dataset ($F$) with rank data, $k$;  
**Output**: Set of clusters ($C$);

1. randomly select $k$ data as initial cluster centers;
2. for every $f_i \in F$
3. (re) calculate distance from all the cluster centers $c_j$, $1 \leq j \leq k$;
4. assign the $f_i$ with the cluster number to its nearest cluster center;
5. 6. calculate mode to identify the updated cluster centers;
7. repeat step 2 to 6 until the convergence criterion is satisfied;
8. return $C$;

1) **EXPLANATION OF KRanks ALGORITHM**

In Line 1 (Algorithm 1), randomly select $k$ numbers of data ($c_j$, $1 \leq j \leq k$, the initial cluster centers) from $F$. For proposed KR2, we use the technique as discussed in section III-A, i.e., select the initial cluster centers in such a manner that the initial selected cluster centers are positioned far apart to each other. Each consumer data $f_i \in F$ distances are calculated $\text{dist}_M(f_i, c_j)$, $1 \leq i \leq n$, from all the clusters centers $c_j$, $1 \leq j \leq k$ (Line 3). For the ranked data, we may apply any one of the four different types ($M$ $\in$ Kendall’s tau, Spearman’s footrule, Spearman’s rho square distance, and Cayley’s) of distance calculations. Line 4, the $f_i \in F$ is assigned the cluster number to its nearest cluster center $c_j$, $1 \leq j \leq k$ to (re) form the clusters. Line 6, calculate mode (cluster center) values $c_j$, $1 \leq j \leq k$ from the newly formed clusters $c_j$, $1 \leq j \leq k$ where the centers consist of ordinal rank values. If the content of the calculated center (mode) is not distinct (i.e., repetitions of the ranks are present in the cluster centers), then apply simple greedy method (used in KR3) to create the center with non repeated ranks (refer next section III-C2). Line 7, repeat steps 2 to 6 till the convergence criterion is satisfied.

2) **FORMATION OF THE CLUSTER CENTER (MODE) WITH NON REPEATED RANKS**

If we calculate simple mode (to capture highest frequency) value to get cluster center then there may be a chance that the
Table 3: The frequency table.

|   | \(f_{11}\) | \(f_{21}\) | \(f_{31}\) | \(f_{41}\) |
|---|---|---|---|---|
| 1 | 6\(_{(p_{11})}\) | 0 | 0 | 0 |
| 2 | 0 | 4\(_{(p_{22})}\) | 0 | 2 |
| 3 | 0 | 0 | 2\(_{(p_{33})}\) | 2 |
| 4 | 0 | 0 | 4\(_{(p_{44})}\) | 2 |

Cluster center contains repeated rank which may affect the cluster quality or may lead to difficulties in campaign selection problem. Hence, to keep the ranks of the cluster centers distinct we should apply the greedy method as follows: Let cluster \(C_i\), \(1 \leq i \leq k\) contains \(q_i\) data, first of all create a frequency table of size \(d \times d\), the entry \(p_{ij}\), \(1 \leq i, j \leq d\), contains the frequency of the \(p\)th rank on \(p\)th column (See example 2, Table 3). Store the frequencies from the frequency table in the variable \(rv\) of size \(1 \times d^2\) from the frequency table row wise. Arrange the content of the \(rv\) in descending order using stable sort algorithm. Next, select the data one by one (from left i.e., maximum frequency first) from \(rv\). Let the first value is \(rv_1\), then identify the column (from left) of the frequency table, where this \(rv_1\) value is present. Assume the table entry \(p_{gh}\), \(1 \leq g, h \leq d\) is matched with the \(rv_1\), i.e., \(g\)th row and \(h\)th column of the frequency table. The value of the column \(h\) or mode[\(h\)] of the mode is set to \(g\) i.e.,

\[\text{mode}[h] = g.\]

In other words, mode is \(\{\_, \_, \ldots, g, \_, \_, \ldots\}\), where index of the \(g\) is \(h\). Lock the column \(h\) and row \(g\) of the frequency table to block further processing, so that no other column can set the same row value. If two or more rows or columns contain the same maximum frequency value \(rv_i\), then select the left column (i.e., if the column contains same maximum value) or select the upper row (i.e., if the column contains same frequency value) to set the column value of the cluster center (mode) with the selected row number. Continue the same process for the rest of the column to form the mode (cluster center, \(C_i\)) which contains the distinct rank values.

Example 2: Formation of the cluster center (mode) with non repeated ranks. Let a cluster contains six data (\(q_i = 6\)) where the dimension is 4 (\(d\)) (Table 2). Last row of the Table 2 exhibits the cluster center (mode value), where rank 2 is repeated two times, hence we have to apply our proposed center identification method with non repeated ranks. A 4 by 4 frequency table (Table 3) is formed. The \(p_{11}\) is 6, since the first column \(I_1\) contains six 1s (rank 1). Next in \(I_2\), four 2s (rank 2) and two 3s (rank 3) are there, hence the entries of \(p_{22}\) and \(p_{33}\) are four and two respectively. Similarly, identify the entry values (frequency) of other columns. Next collect the frequencies from the frequency table row wise to form the \(rv\) vector. After applying the stable sort, \(rv\) contains the data \{6, 4, 4, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0\}. Next scan the frequency table row wise to identify the \(d\) maximum value present in the frequency table. The first data \(rv_1 = 6\), and it is matched with \(p_{11}\) of Table 3 and \(g = 1, h = 1\). Therefore, the first column of the mode (since, \(h = 1\) value is set to 1 (since \(g = 1\)) i.e., mode = \(\{1, \_, \_, \_\}\), and the 1\(^{st}\) row and 1\(^{st}\) column of the frequency table are blocked so that no other column can set the same 1 (rank 1) row value (to avoid repetition of the mode value). Next data \(rv_2 = 4\) is matched with \(p_{22}\), hence the 2\(^{nd}\) column of the mode value is set to 2 (row 2), and the updated mode value is: mode = \(\{1, 2, \_, \_\}\), and the 2\(^{nd}\) row and 2\(^{nd}\) column of the frequency table are blocked. Next data \(rv_3 = 4\), is matched with \(p_{33}\) (4th row and 3rd column, i.e., \(g = 4, h = 3\)) only (since the 2nd row and 2nd column of the frequency table have been blocked already), hence the mode value is: mode = \(\{1, 2, 4, \_\}\). Finally, \(rv_4\) is 2 is matched with \(p_{44}\), 3\(^{rd}\) row only (since, 1st, 2nd and 4th rows, and 1st, 2nd, and 3rd columns have already been blocked), hence the mode value is: mode = \(\{1, 2, 4, 3\}\) (last row of Table 2) which contains distinct rank values.

Time Complexity: The time complexity of the KRanks algorithm is \(O(nktdist)\), where \(n\) is the number of data, \(k\) is the numbers of clusters and \(tdist\) is the time complexity to calculate the distance and \(y\) is the number of iterations (to satisfy the convergence criterion).

D. CAMPAIGN SELECTION PROCESS

In this section we discuss about the campaign selection process using the clusters which have been identified in the previous section. Let \(CG = \{cg_1, cg_2, \ldots, cg_s\}\) be the set of campaigns, where \(s\) is the total number of campaigns, the \(cg_i = \{i_1, i_2, \ldots, i_m\}\), where \(m\) is the number of items present in the \(i\)th campaign and \(i_j \in \Theta\). We have \(k\) number of consumers groups (clusters) \(C_i\), \(1 \leq i \leq k\), where \(c_i\), \(1 \leq i \leq k\) is the center of the cluster. Each \(c_i\), \(1 \leq i \leq k\) is a row vector of size \(1 \times d\). In the process of campaign selection, we can perform two basic tasks as follows: (1) identify items for the forthcoming campaigns for the given groups, (2) identify groups for the given campaigns.

1) IDENTIFICATION OF ITEMS FOR THE FORTHCOMING CAMPAIGNS

Consider a cluster \(C_i\), \(1 \leq i \leq k\), where \(c_i\), \(1 \leq i \leq k\) is the center of the cluster. The center \(c_i\) contains central ranks of the items \(i_j\), \(1 \leq j \leq d\) for the consumer group \(C_i\). For the campaign design, we can select \(k_1\), \(1 \leq k_1 \leq m_i\) items of least ranks values from the \(c_i\), \(1 \leq i \leq k\) for the given cluster \(C_i\) (consumer segment). Note that rank 1 is the most preferred item and rank \(d\) is the least preferred item. The intuition behind the selection of the least ranked items from the cluster center \(c_i\) is, the center \(c_i\) summarizes the preferences on items for the given group \(C_i\). If we can keep the popular items (low ranked) in the campaign then the campaign can draw more attention to the consumers of the selected consumer group.

Next, we present an algorithm for campaign selection:

Using Algorithm 2, set of campaigns can be identified using the center of the clusters of the customers. From each cluster centers \(c_i\), \(1 \leq i \leq k\), \(k\) number of items are identified to form the campaign \(cg_i\). Next we present an example to illustrate the algorithm.

Example 3: Let the cluster center \(c_1 = \{2, 4, 5, 6, 1, 3\}\), hence the ranks of the items \(\{I_1, I_2, I_3, I_4, I_5, I_6\}\) are \(\{2, 4, 5, 6, 1, 3\}\) respectively, here rank 1 is the highest preferred
Algorithm 2 Selection of Campaigns for the Given Cluster

Input: cluster centers, $c_i$, $1 \leq i \leq k$, and $k$;  
Output: set of campaigns ($c_{gi}$) $\in$ CG;  

1. for every $c_i$, $1 \leq i \leq k$  
2. allocate desired $k$ least ranks items from the cluster center $c_i$, $1 \leq i \leq k$, to form the campaign $c_{gi}$;  
3.  
4. return set of campaigns CG;  

item (5th item) (i.e., I5) and rank 6 is the least preferred item (i.e., I4). The least three ranks are {2, 1, 3} for the items {I1, I5, I6} respectively. If $c_{gi}$ is for the cluster $c_i$, 1 $\leq$ $i$ $\leq$ k, is appropriate for the cluster $c_i$, if $\sum R_{ij}$, 1 $\leq$ $j$ $\leq$ k, is minimum. Assume $c_j$ contains the data $|I_{il}|$, $I_{il}$, ..., $I_{im}$ $\in$ $\theta$, where the ranks are $\{\pi_{il}, \pi_{l2}, ..., \pi_{ljm}\}$ respectively. The $\pi_{ij}$ is the rank of the item $I_{ij}$. 

Definition 4: Suitability of a campaign $c_{gi} = \{I_{i1}, I_{i2}, ..., I_{im}\}$ with respect to cluster center $c_j$ of $C_j$ is identified by lowering $\sum R_{ij}$. The sum $\pi_{ij}$ value higher the suitability of the campaign i.e., the $i^{th}$ campaign is suitable for $j^{th}$ cluster (consumer group), if $\sum R_{ij}$ is lower.  

Lemma 1: If the rank value $\pi_{ij}$ of $l^{th}$ attribute $I_{il}$, 1 $\leq$ $l$ $\leq$ d in the cluster center $c_i$, 1 $\leq$ $i$ $\leq$ k, is more than the rank value $\pi_{ij}$ of the same $l^{th}$ attribute in the cluster center $c_j$, 1 $\leq$ $j$ $\leq$ k, where $I_{ij} = I_{il}$ $\cap$ $c_i$ $\cap$ $\theta$, i $\neq$ j (i.e., $\pi_{ij} > \pi_{il}$), then the $l^{th}$ attribute is more popular in group $C_j$ than group $C_i$.  

Proof: To form the cluster center, we measure the frequency of the values of the attributes and the highest frequency value is set as the rank value of the cluster center for the attribute. Moreover, as we know that less rank value indicates more popular item. Hence, if the attribute $I_{il}$ contains less rank value $\pi_{il}$ in the cluster center $c_i$, than the rank value $\pi_{ij}$ in the cluster center $c_j$, we can say that the attribute $I_{ij}$ is more popular in cluster group $C_j$ than $C_i$, since more people from cluster group $C_j$ likes (giving less rank) than $C_i$.  

Theorem 1: Let $c_{gi}$ be a campaign with size $|c_{gi}|$, and $c_i$ and $c_j$ be the cluster centers of the clusters $C_i$ and $C_j$ respectively. If $\pi_{il} < \pi_{ij}$, $I_{il} \in c_{gi}$ $\cap$ $c_i$, $I_{il} \notin c_{gi}$ $\cap$ $c_j$, then $\sum R_{il} < \sum R_{ij}$.  

Proof: The $c_i$ and $c_j$ be two cluster centers of $C_i$ and $C_j$ 1 $\leq$ $i$, $j$ $\leq$ k. The ranks $\pi_{il}$ of $I_{il}$ $\in$ $c_{gi}$ $\cap$ $c_i$ and $\pi_{ij}$ of $I_{il}$ $\in$ $c_{gi}$ $\cap$ $c_j$ are from the clusters $C_i$ and $C_j$ respectively.

If attribute $I_{il}$ is more popular in $C_i$ than $C_j$, then from Lemma 1 i it is proved that $\pi_{il} < \pi_{ij}$. Accordingly, $\sum \pi_{il} \in c_i \cap c_{gi} < \sum \pi_{ij} \in c_j \cap c_{gi}$, $\sum R_{il} < \sum R_{ij}$. This proves the theorem.  

Example 4: Let the three campaigns $c_{g1} = \{I_{11}, I_{12}, I_{13}\}$, $c_{g2} = \{I_{24}, I_{25}, I_{26}\}$, and $c_{g3} = \{I_{31}, I_{33}, I_{35}, I_{36}\}$ and we have three clusters (which we have already formed using Section III.C) where three clusters centers $c_1 = \{I_{11}, I_{12}, I_{13}, I_{14}, I_{15}, I_{16}\}$ and the ranks are {1, 2, 3, 4, 5, 6} respectively. $c_2 = \{I_{11}, I_{12}, I_{23}, I_{24}, I_{25}, I_{26}\}$ and the ranks are {6, 5, 4, 3, 2, 1} respectively, similarly $c_3 = \{I_{31}, I_{32}, I_{33}, I_{34}, I_{35}, I_{36}\}$ and the ranks are {3, 1, 4, 2, 6, 5}. The values of the cluster centers exhibits the rank of the items, e.g., the values {3, 1, 4, 2, 6, 5} of $c_3$ indicates that the rank of the 1st item ($I_{31}$) is (31), the rank of the 2nd item ($I_{32}$) is (32) and so on. The values of the evaluation parameter for the campaign $c_{g1}$ (for 1st, 2nd and 3rd items) are $\sum R_{i1} = (1 + 2 + 3)/3 = 2$, $\sum R_{i2} = (4 + 5 + 6)/3 = 5$ and $\sum R_{i3} = (3 + 1 + 4)/3 = 2.66$ using cluster centers $c_1, c_2$, and $c_3$ respectively. Since for the campaign $c_{g1}$, the $\sum R_{i1}$ value is the smallest, hence $c_{g1}$ campaign is appropriate for cluster 1 (or consumer group 1). Similarly, for $c_{g2}$ (for items 4th, 5th and 6th), the values of the evaluation parameters are $\sum R_{i2} = (4 + 5 + 6)/3 = 5$, $\sum R_{i3} = (3 + 2 + 1)/3 = 2$ and $\sum R_{i3} = (2 + 6 + 5)/3 = 3.33$, since for the campaign $c_{g2}$, the $\sum R_{i2}$ value is the smallest, hence $c_{g2}$ campaign is appropriate for cluster 2 (or consumer group 2). Similarly, $\sum R_{i1} = (1 + 3 + 5 + 6)/4 = 3.75$, $\sum R_{i2} = (6 + 4 + 2 + 1)/4 = 3.25$ and $\sum R_{i3} = (3 + 4 + 6 + 5)/4 = 4.5$ are the evaluation parameter of the $c_{g3}$ (for the items 1st, 3rd, 5th, and 6th), since for the campaign $c_{g3}$, the $\sum R_{i3}$ value is the smallest, hence $c_{g3}$ campaign is appropriate for cluster 2 (or consumer group 2). For consumer group $c_3$, no existing campaign is appropriate, hence we have to redevelop a new campaign for consumer group $c_3$.  

Now, we present an algorithm for identification of consumer group for the given campaigns:

Algorithm 3 Identification of Clusters (Consumer Group) for the Given Campaigns

Input: set of campaigns $c_{gi}$, 1 $\leq$ $i$ $\leq$ s and set of clusters with cluster centers $c_i$, 1 $\leq$ $j$ $\leq$ k;  
Output: (Campaign - Customer group pair) i.e., $(c_{gi}, c_j)$, 1 $\leq$ $i$ $\leq$ s, 1 $\leq$ $j$ $\leq$ k;  

1. For every campaigns $c_{gi}$ {  
2.  
3. make pair $(c_{gi}, c_j)$ i.e., campaign $c_{gi}$ is appropriate for cluster $c_j$;  
4.  
5. Return $(c_{gi}, c_j)$, 1 $\leq$ $i$ $\leq$ s, 1 $\leq$ $j$ $\leq$ k;  

The time complexity of Algorithm 2 is $O(kd \log d)$ and the time complexity of the Algorithm 3 is $O(skd)$, where $k$ is the
number of clusters, $d$ is the dimension of the data, $s$ is the number of campaigns.

IV. EXPERIMENTS

A. DATASET

To demonstrate the effectiveness of our proposed algorithms on ranked data, we have used two real ranked datasets Sushi [28] and American Psychological Association (APA) presidential election [29]. The numbers of instances are 5000 and 5738 and the dimensions are 10 and 5 respectively. In Sushi data we have up scaled the data by 1 i.e., initial ranking range lies between 0 to 9, up scaled ranking range lies between 1 to 10. Moreover to illustrate the scalability of our proposed algorithms we have executed all the algorithms on different dimensions and sizes where the data (the synthetic data generated by Python program) are purely ranked data.

B. PERFORMANCE EVALUATION PARAMETERS

To evaluate our proposed algorithms we have selected evaluation techniques as follows:

1) EVALUATION PARAMETERS FOR CLUSTERING ALGORITHMS

(a) The overall average Silhouettes width (OASW) [30]: The $S_i = \frac{1}{n} \sum_{y \in C_i} d(y, C_i \backslash \{y\})$, where $d_i \in C_i$ is the average dissimilarity of $f_i$ data to all other data of cluster $C_i$, where the data $f_i$ belongs to cluster $C_i$, $1 \leq i \leq k$, $y$ is the min AvgDist($f_i$, $C_j$) (minimum average distance), $1 \leq j \leq k$, $j \neq i$ i.e., the minimum average distance of $f_i$ data to all data in any other cluster in which $f_i$ does not belong to. Note that $-1 \leq S_i \leq 1$. The Overall average Silhouettes width (OASW) is the average of the $S_i$, $1 \leq i \leq n$, i.e., OASW = $\frac{1}{n} \sum_{i=1}^{n} S_i$, where $n$ is the number of data. Note that more OASW value means better cluster. (b) Validity index (VI) [31]: VI = $\frac{\text{Intra}}{\text{Inter}}$, where $\text{Intra} = \frac{1}{k} \sum_{i=1}^{k} \sum_{j \in C_i} ||f_i - c_i||^2$ and $\text{Inter} = \min \{||c_i - c_j||^2, i, j = 1, 2, \ldots, k \}$, $j \neq i$. Note this validity measure should be as minimum for the best quality of the cluster.

2) EVALUATION PARAMETERS FOR CAMPAIGN SELECTION

(c) Identification of items for the forth coming campaign: First, from the identified cluster centers $c_i$, $1 \leq i \leq k$, for each consumer groups $C_i$, $k$ number of least rank items are identified (from the cluster centers $c_i$) to set the campaigns. For each consumer group $C_i$, we calculate local attention factor \( \text{LAF}_i = \frac{\sum_{j=1}^{k_i} LAF_{ij}}{k_i} \) and then, calculate final attention factor \( \text{FAF} = \frac{\sum_{C_i} \text{LAF}_i}{k} \). From the attention factors our aim is to quantify the attention factor of the items which will attract by the consumer groups. Consumers give lowest rank to their favorite items and thus a low FAF value indicates high consumer attentions. To observe the need of consumer segmentation $\text{GAF}$ (grand attention factor) value is identified. For comparison purpose we may set single campaign for all the $n$ consumers. Single grand center ($c_{\text{grand}}$) is identified for all the consumers and single campaign using $k_1$ least items is defined. The calculated GAF = $\frac{1}{n} \sum_{j=1}^{n} \frac{1}{\text{FAF}}$. If $\text{GAF} \leq \text{FAF}$ then no need to divide the consumers into groups and single campaign is sufficient for marketing. But if $\text{GAF} > \text{FAF}$ then divide the consumers into groups and set the campaign separately for the different consumer groups. (d) Identification of the groups for the given campaigns: First, each campaign $c_g$ is assigned to the respective consumer group by calculating $\text{sumR}_{ij}$ using cluster centers. If $\text{sumR}_{ij}$ is lower then the $i^{th}$ campaign is suitable for $j^{th}$ cluster (consumer group). Now to validate $\text{sumR}_{ij}$, we calculate local attention factor $\text{LAF}_{ij} = \frac{\sum_{c_g} \text{sumR}_{ij}}{|c_g|}$, $1 \leq i \leq k$, $1 \leq j \leq s$. The lower $\text{sumR}_{ij}$ and $\text{LAF}_{ij}$ values show more consumer attention of $i^{th}$ consumer group of the $j^{th}$ campaign. In summary, $\text{LAF}_{ij}$ is used to validate $\text{sumR}_{ij}$ values i.e., cluster center is sufficient to identify the consumer group of the given campaign.

C. EXPERIMENTAL ANALYSIS

The KR1, KR2 and KR3 are the proposed algorithms. In KR1 algorithm, the initial cluster centers are selected randomly. In KR2, the initial cluster centers are selected based on our proposed technique (Sec. III-A). In both KR1 and KR2, simple mode is calculated to update the cluster centers. In KR3, the initial clusters centers are selected based on (Sec. III-A). If any center contains repeated rank then our proposed technique (Sec. III-C2) is applied. For comparison purpose, we have implemented simple K-Means [32] (KM), K-Medoids [33] (KMed), and K-Mode [34] (KMo). For distance calculation we have used (for KR1, KR2, KM, KMed, and KMode) Spearman’s footrule distance (footDist), Spearman’s rho square distance (smDist), Kendall’s tau distance (kendist), and Cayley distance (cayley). In identification of items for the forth coming campaign we use $k = 3$. Two sets of items $c_{g1} = \{2, 4, 5, 7\}$, $c_{g2} = \{1, 3, 6, 7\}$ for Sushi dataset and $c_{g3} = \{1, 3, 4\}$, $c_{g4} = \{1, 2, 4\}$ for APA dataset are considered for identification of the groups for the given campaigns respectively. To evaluate the campaign selection process, in KM all the final cluster centers are converted into nearest integer value. All the algorithms iterate until the total mean difference between current and previous means is less than 0.001 (we set this value as convergence criteria). The algorithm has been executed to generate three to five clusters. The values of the evaluation parameters have been reported for the same. To show the execution efficiency, we have executed all the algorithms on different dimensions for the fixed number of data (5000) and cluster (3). Moreover, we have executed all the algorithms on different data size for the fixed number of dimensions (10) and cluster (3). All the experiments were conducted on a PC with hardware Intel Core(TM) i5-4570 processor, (3.20 GHz), and 8GB RAM running on Windows 10. Whereas, on the software end with Python 3.6 running on eclipse neon IDE.
V. RESULT AND ANALYSIS

Now we discuss the results and analysis of all the algorithms. In Table 4 (using Sushi dataset), the overall average Silhouettes width calculations (OASW) for all distance measure techniques are displayed. We may recall that Algorithms KR1, KR2, KR3 are our proposed algorithms, with distance calculations as in Table 4, most of the time the KR3 and KR2 performed well to form the cluster (except Table 4b, where KR1 performed well when number of clusters was 3). In Table 5, the validity index values are displayed. The results demonstrated that most of the time KR3 technique performs best (except Table 5b and 5c where we get a mixed response as KR1 and KR3 perform well).

The performance on APA data where OASW and validity indices values are exhibited in Tables 6, 7 respectively. The results demonstrated that most of the time KR2 and KR3 give best performance (except Table 6c, Table 6d where KR1 performs best for cluster number 3, 5 respectively). Table 7, displayed the mixed responses among techniques. KR3 displayed great results in Table 7c. Whereas in 7a, 7b, 7d KR1, KR2 and KR3 gave best performance for different cluster numbers.

After applying all the distance measures for ranked clustering we reached to the conclusion based on the result (Table 4 to Table 7) of evaluation parameters that Kendall’s tau distance measure produces best results. So, the following...
experiments on campaign selection are performed using Kendall’s tau distance measure only.

The convergence graph have been presented in Figures 1 and 2 for Sushi and APA dataset respectively for $k = 3$ using Kendall’s tau distance. We have observed that after few iterations the total mean difference is near zero. From the plot the convergence criteria can be set. To show the execution efficiency, we have executed all the algorithms on different dimensions (maximum 25 dimension) for the fixed number of data (5000) and cluster (3) shown in Figure 3 using all the distance measures. Moreover, we have executed all the algorithms on different data sizes (maximum data 7000) for the fixed number of dimensions (10) and cluster (3) shown in Figure 4 using all the distance measures. We have observed

### TABLE 8. Identification of items for the forth coming campaign.

| Dataset | Sushi | APA |
|---------|-------|-----|
| Techniques | $F_A F$ | $G_A F$ | $F_A G$ | $G_A F$ |
| KR1     | 3.1670 | 4.0540 | 2.4262 | 2.9150 |
| KR2     | 4.1762 | 5.0540 | 2.3578 | 2.9150 |
| KR3     | 4.2708 | 5.0540 | 2.3764 | 2.9150 |
| KM      | 4.1900 | 5.0540 | 2.3891 | 2.9150 |
| KMo     | 4.2282 | 5.0540 | 2.3962 | 2.9150 |
| KMe     | 5.2145 | 5.2540 | 2.7340 | 2.9150 |

FIGURE 1. Convergence plot for Sushi dataset.

FIGURE 2. Convergence plot for APA dataset.

FIGURE 3. Dimension versus time.

FIGURE 4. Execution efficiency comparison.
that our proposed algorithms (KR1, KR2 and KR3) are time efficient as compared to existing clustering techniques.

**Experiments on Campaign Selection:** Table 8 demonstrates the experiments on identification of items for the forthcoming campaign. In almost all the cases the FAF values are smaller than the GAF value, and the lower FAF value than the GAF value shows the applicability of the consumer segmentation. If we consider only the FAF values of the Sushi data, we have seen that KR1 and KR2 values are lower than the rest of the data. In APA data the FAF values of KR2 and KR3 are less than the rest of the data. Experimental results of identification of the groups for the given campaigns on Sushi Data (Table 9) and APA Data (Table 10) are as follows. Almost all the cases (except in Table 10 in KM technique using campaign items {1, 3, 4}) the LAF and sumR values show the lowest in the same cluster number. Hence, results show the applicability of the use of the sumR values for consumer segment selection process for the given campaign.

**VI. CONCLUSION**

In this paper first we group the ranked data of consumer feedback, by applying different distance calculations, e.g., Kendall’s tau, Spearman’s rho square, Spearman’s footrule, Cayley’s distance and then use the formed groups in campaign selection process. We have compared our proposed clustering algorithm with existing algorithms on different real datasets, and results showed the effectiveness of our proposed algorithm.

**REFERENCES**

[1] P. Kotler and K. Keller, *Marketing Management*, 15th ed. London, U.K.: Pearson, 2015.
[2] E. Brentari, L. Dancelli, and M. Manisera, “Clustering ranking data in market segmentation: A case study on the Italian McDonald’s customers’ preferences,” *J. Appl. Statist.*, vol. 43, no. 11, pp. 1959–1976, Aug. 2016.
[3] D. J. Best and J. C. W. Rayner, “Analysis of ranked data in randomized blocks when there are missing values,” *J. Appl. Statist.*, vol. 44, no. 1, pp. 16–23, Jan. 2017.
[4] M. Alvo and P. Cabilio, “Rank correlation methods for missing data,” *Can. J. Statist.*, vol. 23, no. 4, pp. 345–358, Dec. 1995.
[5] J. Jacques and C. Biernacki, “Model-based clustering for multivariate partial ranking data,” *J. Stat. Planning Inference*, vol. 149, pp. 201–217, Jun. 2014.
[6] L. M. Busse, P. Orbanz, and J. M. Buhmann, “Cluster analysis of heterogeneous rank data,” in *Proc. 24th Int. Conf. Mach. Learn. (ICML)*, Corvallis, OR, USA, Jan. 2007, pp. 113–120.
[7] R. M. Ursu, “The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions,” *Marketing Sci.*, vol. 37, no. 4, pp. 530–552, Aug. 2018.
[8] P. Diaconis and R. L. Graham, “Spearman’s footrule as a measure of disarray,” *J. Roy. Stat. Soc. B, Methodol.*, vol. 39, no. 2, pp. 262–268, Jan. 1977.
[9] M. G. Kendall, “A new measure of rank correlation,” *Biometrika*, vol. 30, nos. 1–2, pp. 81–93, Jun. 1938.
[10] J. I. Marden, *Analyzing and Modeling Rank Data*. Boca Raton, FL, USA: CRC Press, 2014.
[14] P. Werrij and R. Kaptein, “Clustering ordinal survey data in a highly structured ranking,” Vrije Univ. Amsterdam, Amsterdam, The Netherlands, Res. Paper ACM ISBN 0-12345-67-8/9/001, 2016.

[15] M. Grbovic, N. Djuric, S. Guo, and S. Vucetic, “Supervised clustering of label ranking data using label preference information,” Mach. Learn., vol. 93, nos. 2–3, pp. 191–225, Nov. 2013.

[16] A. Taristano, “Weighted rank correlation and hierarchical clustering,” in Proc. Book Short Papers CLADAG, Parma, 2005, pp. 517–521.

[17] P. Viappiani, “Characterization of scoring rules with distances: Application to the clustering of rankings,” in Proc. 24th Int. Joint Conf. Artif. Intell. (IJCAI), 2015, pp. 104–110.

[18] C. Chan, “Intelligent value-based segmentation method for campaign management: A case study of automobile retailer,” Expert Syst. Appl., vol. 34, no. 4, pp. 2754–2762, May 2008.

[19] J. Paetz, “Campaign management design based on segmentation by rank clusters,” J. Marketing Anal., vol. 3, no. 4, pp. 187–214, Dec. 2015.

[20] P. J. Fleming and M. A. Pashkevich, “Optimal advertising campaign generation for multiple brands using MOGA,” IEEE Trans. Syst., Man, Cybern. C. Appl. Rev., vol. 37, no. 6, pp. 1190–1201, Nov. 2007.

[21] R. Michel, I. Schnakenburg, and T. von Martens, “Effective customer selection for marketing campaigns based on net scores,” J. Res. Interact. Marketing, vol. 11, no. 1, pp. 2–15, Mar. 2017.

[22] X. Peng, H. Zhu, J. Feng, C. Shen, H. Zhang, and J. T. Zhou, “Deep clustering with sample-assignment invariance prior,” IEEE Trans. Neural Netw. Learn. Syst., early access, Dec. 31, 2020, doi: 10.1109/TNNLS.2019.2958324.

[23] X. Peng, Z. Huang, J. Lv, H. Zhu, and J. T. Zhou, “Comic: Multi-view clustering without parameter selection,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 5092–5101.

[24] X. Peng, J. Feng, S. Xiao, W.-Y. Yau, J. T. Zhou, and S. Yang, “Structured AutoEncoders for subspace clustering,” IEEE Trans. Image Process., vol. 27, no. 10, pp. 5076–5086, Oct. 2018.

[25] P. S. Bishnu and V. Bhattacherjee, “Software fault prediction using quad tree-based K-Means clustering algorithm,” IEEE Trans. Knowl. Data Eng., vol. 24, no. 6, pp. 1146–1150, Jun. 2012.

[26] S. J. Redmond and C. Heneghan, “A method for initialising the K-means clustering algorithm using KD-trees,” Pattern Recognit. Lett., vol. 28, no. 8, pp. 965–973, Jun. 2007.

[27] B. Mirkin, “Clustering for data mining: A data recovery approach,” Dept. Comput. Sci. Inf. Syst., Birkbeck Univ. London, London, U.K., Tech. Rep., 2005.

[28] T. Kamishima, “Nantonnac collaborative filtering: Recommendation based on order responses,” in Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining (KDD), 2003, pp. 583–588.

[29] P. Diaconis, Group Representations in Probability and Statistics (Lecture Notes-Monograph Series), vol. 11. Ann Arbor, MI, USA: Institute of Mathematical Statistics, 1988, p. 1–192.

[30] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” J. Comput. Appl. Math., vol. 20, pp. 53–65, Nov. 1987.

[31] S. Ray and R. H. Turi, “Determination of number of clusters in k-means clustering and application in colour image segmentation,” in Proc. 4th Int. Conf. Adv. Pattern Recognit. Digit. Techn., Kolkata, India, 1999, pp. 137–143.

[32] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in Proc. 5th Berkeley Symp. Math. Statist. Probab., Oakland, CA, USA, vol. 1, 1967, pp. 281–297.

[33] L. Kaufman and P. Rousseeuw, Clustering By Means of Medoids, Amsterdam, The Netherlands: North Holland, 1987.

[34] Z. Huang, “Extensions to the K-means algorithm for clustering large data sets with categorical values,” Data Mining Knowl. Discovery, vol. 2, no. 3, pp. 283–304, Sep. 1998.