Survey Research Methods (2021)
Vol. 15, No. 2, pp. 169-185
doi:10.18148/srm/2021.v15i2.7773
European Survey Research Association

Directional Pattern based Clustering for Quantitative Survey Data: Method and Application

Roopam Sadh
School of Computer & Systems Sciences
Jawaharlal Nehru University
New Delhi, India

Rajeev Kumar
School of Computer & Systems Sciences
Jawaharlal Nehru University
New Delhi, India

The paper proposes pattern based clustering technique for survey data that uses the convention of direction instead of magnitude. Instead of arbitrary manual settings of clustering parameters, the clustering operates upon theoretically justifyable groupings in the data. The resulting clusters are then automatically linked to those groupings. The paper applies the proposed method to real world survey data on the relevance of quality performance indicators in higher education. The results are compared to the corresponding results of *K*-Means clustering. The results suggest that the proposed pattern clustering method performs better in segregating respondents according to stakeholder theory making the clusters more meaningful. These results are taken as evidance for the suitability of pattern based analysis methods for quantitative survey data.

Keywords: Pattern clustering; Survey data; Higher education; Institutional quality; Stakeholder theory

1 Introduction

Surveys are the most widely used data collection method in organizational and behavioral research (Behrend, Sharek, Meade, & Wiebe, 2011; Huang, Liu, & Bowling, 2015). Some of the key domains that utilize surveys frequently are policy making, higher education, health care, psychology, and market research (Church & Waclawski, 2001; Ward & Meade, 2018). Organizations conduct surveys to obtain better understanding of perceptions, interests, and behavior of their stakeholders (Church & Waclawski, 2001). Due to wide range of applications, proper analysis of survey data becomes a crucial concern. Improper analysis of survey data lead to meaningless and sometimes weird findings. Selection of appropriate analytical tools is therefore a prerequisite for achieving meaningful insights from survey data (Vandervalk, Louch, Guerre, & Margiotta, 2014). However, the reliability of analysis tools largely depends on the type of data and the purpose of the application (Estivill-Castro, 2002; Rodriguez et al., 2019).

Quantitative survey data possesses several distinct characteristics (Sadh & Kumar, 2020). It contains data of the same type and fixed value-range for almost all of its dimensions i.e., fixed ordinal marking scale (Lee, Jones, Mineyama, & Zhang, 2002). Respondent category labels are also valuable as relationships among variables with respect to these categories are generally sought (Sadh & Kumar, 2020). Further, marking-patterns in surveys are of utmost importance since they reflect respondents' behavior (Grice, 2015). Due to these characteristics, survey data require dedicated analytical techniques. We are particularly interested in clustering of survey data as it has some significant applications e.g., identification of distinct behaviors, classification, pattern finding, validation of theories etc., (Tan, Steinbach, Kumar, et al., 2006).

Clustering is used to divide the objects into groups in such a way that the objects in same group are similar but dissimilar to the objects in other groups (Guha, Rastogi, & Shim, 2000). Since, several clustering methods are available each having its own features and limitations (Xu & Tian, 2015) hence one has to choose a method according to the purpose and type of data (Estivill-Castro, 2002; Rodriguez et al., 2019). Most of the existing clustering methods generally define closeness in terms of value-based similarity (Kriegel, Kröger, & Zimek, 2009; H. Wang & Pei, 2008) whereas survey data is more appropriate for pattern-based similarity since patterns of marking in survey data reflect the behavior of the respondents (Grice, 2015; Valsiner, Molenaar, Lyra, & Chaudhary, 2009).

A few pattern based clustering methods exist in the literature. They come under the category of sub-space clustering and are designed especially to cluster high dimensional data of a specific kind, e.g., DNA micro array data (Jiang, Tang, & Zhang, 2004; Kriegel et al., 2009). In contrast, survey data has a much smaller dimensional space of global nature and it contains small ordinal values for all of its dimensions (Lee

Contact information: Roopam Sadh, Data to Knowledge (D2K) Lab, Room No. 221, SC&SS, JNU, New Delhi-110067, India. (E-mail: roopam.sadh@gmail.com)

et al., 2002). Thus, existing pattern clustering methods are not applicable for survey data. Further, survey data contains side information in the form of category labels which can be used to obtain superior results and to decide on clustering parameters e.g., number of clusters (Wagstaff, Cardie, Rogers, Schrödl, et al., 2001; Xing, Jordan, Russell, & Ng, 2003). Due to such reasons, survey data requires a specialized pattern clustering method.

This paper proposes a pattern based clustering method designed for quantitative survey data containing ordinal marking values. We name the method *Pattern Clustering for Survey Data based on Directional Differences* (PCSD3). The proposed method utilizes the convention of direction instead of magnitude. It treats each survey observation as a vector of directions rather than a series of values and detects the directional difference between variables. The method explores the dominant patterns of the directional differences through an adaptive decision making procedure. This adaptive procedure uses frequency based probabilistic scoring designed specifically for small ordinal values. Further, PCSD3 does not require manual setting of clustering parameters (number of clusters) since it automatically detects respondent categories, identifies their representative features, and clusters the data accordingly. We therefore claim that the method produces more meaningful and interpretable results according to the properties of survey data.

We apply PCSD3 and K-means clustering using survey data that contains the opinions of academic stakeholders regarding various qualities of higher educational institutions (HEIs). We compare the results of both PCSD3 and K-means with respect to interpretability and usability. For verifying the results we use well established stakeholder theory (A. Burrows & Harvey, 1992; Vroeijenstijn, 2003). Results show that each PCSD3 cluster contains a fair majority of responses from a particular category. This implies that PCSD3 segregates survey responses according to the natural stakeholder grouping. However, this is not true in case of K-means. Hence, the results suggest that PCSD3 is more suitable for quantitative survey data. Further, PCSD3 labels clusters with the names of respondent categories which makes PCSD3 clusters easy to interpret.

Overall, the results of PCSD3, and its comparison with K-means suggest that PCSD3 performs better for quantitative survey data containing ordinal values. Results also empirically validate earlier studies that advocated the use of pattern based measures for behavioral studies (Grice, 2014; Manicas, 2006). The proposed method is useful for several applications. It can be used as a tool to analyze the appropriateness of different stakeholder groupings and to identify general tendencies of groups. The increasingly complex world requires more flexible modeling techniques (Hill et al., 2019; Kern, Klausch, & Kreuter, 2019). In that sense, the proposed method paves the way for developing more sophisticated pattern clustering methods by adapting it with the machine learning paradigm. We envisage that future variants of PCSD3 will remove its limitations and will be able to deal with data containing values of nominal, ordinal, metric, and mixed types.

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

- The study proposes a pattern clustering method specially designed for quantitative survey data containing ordinal values.
- The study empirically validates earlier studies advocating the use of pattern based analysis approaches for behavioral studies.

The paper is organized as follows: Section 2 presents an example scenario and related work that motivated the study. PCSD3 is introduced in Section 3. Section 4 gives a brief introduction to stakeholder theory and Section 5 describes the survey data used. Section 6 uses the data to compare the results of PCSD3 with K-means, and discusses the features and limitations of PCSD3. Finally, Section 7 concludes the paper.

2 Motivation

2.1 Motivating Example

In surveys, 5 to 7 Likert levels (Lee et al., 2002) are often used as indicator for respondents' preferences. We claim that magnitude based treatment of these ordinal quantities infer no meaningful information (Sadh & Kumar, 2020) whereas pattern of marking tells much more about the opinions of respondents i.e., two respondents can be said to share similar opinions, if their marking patterns are similar though their marking values may vary. For clarifying this phenomenon we depict example data with seven observations from three respondent categories (A, B, C) along with the trends of marking in Figure 1. We assume four variables (V1, V2, V3, V4) and a five-level marking scale. The data will be used as running example throughout.

Applying K-means ($K = 3$) on the example data results to observations 1, 3, 4, and 5 ending up in the first cluster, 6 and 7 in second, and 2 alone in the third. However, if we give a close look over the trends of observations then we find that patterns of observations 1, 4, and 7 are identical but their marking values are different. Here pattern similarity can be understood as the relative significance of variables e.g., first and second variables are of low but equal significance, third variable is of highest significance, and fourth have slightly lower significance. In spite of the perfect match in their patterns these observations are clustered separately by K-means. We have also tried different values of K (2, 3, 4, 5) and found that each time observation 7 stood separately from observation 1 and 4. Since K-means is sensitive to magnitudes (value

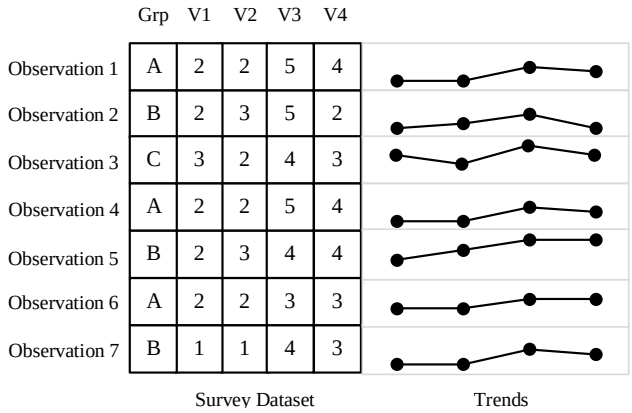


Figure 1. Example survey dataset

based similarity), it suppresses delicate patterns (trends), if the dimensions differ in terms of magnitude. The same is also true with hierarchical clustering. We have applied single linkage (agglomerative) clustering on the example data and found that it also treats observation 7 quite differently than observation 1 and 4.

Similarly, observation 2, shows more pattern-wise similarity with observation 5. However they lie in different clusters for both K-means as well as single linkage. These example thus suggests that for ordinal scaled variables, pattern based clustering is more suitable than value based clustering.

2.2 Existing Pattern Clustering Methods

Existing pattern clustering methods are designed to cluster high dimensional data for specific purposes. High dimensional data poses several challenges for clustering algorithms. One of the major challenges is the presence of irrelevant dimensions. Second, different subsets of dimensions in such data may have correlations. Further, the different subsets of dimensions may relate to different clusters (Kriegel et al., 2009). Thus, clustering of high dimensional data seek similarity between objects with respect to the different subsets of the dimensions instead of the full dimensional space. Sub-space clustering is a widely used term for this type of clustering approach.

Since existing pattern clustering methods treat the data space and the data objects interchangeably, they are called biclustering or two-mode clustering methods (Van Mechelen, Bock, & De Boeck, 2004). The applications that require partitioning of dimensions with respect to similarity in objects and vice-versa demand such a type of clustering approach. Clustering of DNA micro array data is an example application which induced the development of pattern based clustering. Cheng and Church (2000) were the first to introduce the bicluster model. They developed the mean squared residue based node deletion algorithm that simultaneously clusters both, genes (objects) and conditions (dimensions). For making the bicluster model more general Yang, Wang, Wang, and Yu (2002) proposed δ-clustering which allows participation of empty attribute values and devised a move-based algorithm to produce near optimal results. The p-cluster model (H. Wang, Wang, Yang, & Yu, 2002) was the next improvement that added determinism into δ-clustering by finding all qualified biclusters. The p-cluster model was the first bicluster model that defined the similarity on the basis of patterns. To speed up the clustering operation and to make the method scalable Pei, Zhang, Cho, Wang, and Yu, 2003 introduced the concept of maximal pattern (MaPle) that avoids redundant clusters in its mining process.

Since the p-cluster model considers strict shifting and scaling patterns, Liu and Wang (2003) proposed the concept of order preserving clustering approach (OP-cluster) to remove these limitations. The approach is able to capture the general tendency of objects across subsets of dimensions. Apart from these advances H. Wang, Chu, Fan, Yu, and Pei, 2004 introduced the concept of sequence based pattern similarity (SeqClus) which opens the scope of biclustering for more bulky and sequential data.

All of the methods mentioned above are of sub-space clustering type that are designed to cluster high dimensional data. In contrast, survey data and its applications are quite different. Survey data contains fewer dimensions and is defined precisely with respect to a particular objective. Thus, similarity inside survey data cannot be defined on the basis of random sub-spaces. In other words, the nature of survey data in the context of the dimensional space is global. Further, survey data contains small ordinal values that are not suitable for existing methods. Therefore, sub-space clustering is not applicable for survey data. This gap motivates us to develop a dedicated pattern clustering method for survey applications.

3 PCSD3: Architecture & Mechanism

The proposed method divides the survey data on behalf of the marking patterns of the respondents. The method converts the data into direction vectors recording the pattern information of respondents' preferences. The method then detects respondent categories with the help of category labels and divides the data to identify the reference vectors (distinguished features) of the respondent categories. To identify reference vectors, frequency based probabilistic scores are utilized. The convention of direction vectors and frequency based probabilistic scores are used since they are more appropriate in dealing with ordinal values. The entire data of direction vectors is finally matched to reference vectors. The matching procedure detects similarity by measuring the directional differences between the direction vector and the reference vector regarding each dimension.

The whole clustering procedure can be divided into three parts:

1. Creation of direction vectors,

Algorithm 1 (Calculation of frequency based probabilistic scores)

Let the frequencies of −1, 0, and 1 for a variable are X, Y, and Z respectively.

1: Find the frequency of most prevalent direction value (largest frequency).
2: Check, whether 0 is most prevalent (Y is largest)
3: **if** 0 is most prevalent **then**
4: Find second largest frequency and subtract least frequency from it (removing out the weightage of least frequency). Resultant quantity is an intermediate value represented by R.
5: Find probability of R with respect to Y, which is the magnitude of final score. Sign of the score will remain same as of the value corresponds to second largest frequency. Symbolically: R = Second largest frequency – Least frequency S = sign (R/(R+Y)) i.e., sign is minus, if X is second largest
6: **else**
7: Subtract from most prevalent frequency, the frequency of opposite extreme e.g. if X is largest than subtract Z from X (removing out the weightage of other extreme). Resultant is R.
8: Find probability of R with respect to Y (If Y is not 0) or second largest frequency (If Y is 0), which is the magnitude of final score. Sign of score will be adopted from most prevalent value. Symbolically: R = largest frequency – frequency of other extreme S = sign (R/(R+I) I = Y if (Y! = 0) else I = Second largest frequency
9: **end if**

2. Identification of reference vectors, and
3. Matching

The Overall architecture of the proposed method is shown in Figure 2. The following subsections describe each of these steps in turn.

3.1 Creation of the direction vectors

In this step, observations are converted into patterns. A pattern is a vector containing a series of direction values (−1, 0, or 1) where each value defines the weight of a survey item with respect to the preceding item. It is −1, if the value of a variable is less than the value of variable preceding it, 0 if both the values are same, and 1 if the value is larger than the value that precedes. The direction value of the first variable is decided by comparing its value with half of the maximum scale ($n/2$ for an odd scale and $(n + 1)/2$ for an even scale, where n denotes marking levels used). In this paper, we call the direction value of a variable the "directional difference" and the series of direction values for an observation as the "direction vector". Figure 3 shows the direction vectors for the example data.

Usually, marking scale ranges from four to seven (Lee et al., 2002) levels. While such ranges are suitable for representing the opinions of respondents they tend to suppress dissimilarity on the aggregate level, e.g., two variables that are different in terms of variability may possess equal mean. Further, magnitude of differences between observations influences overall similarity. For instance, if two observations are almost similar in the majority of dimensions but are highly dissimilar on a single dimension, this single large difference may influence the clustering result. Due to these reasons, we adopt convention of directional difference instead of magnitude difference. As the direction vector only records the differences in variable directions and not magnitudes, small and large differences are treated equally. In this regard PCSD3 defines similarity in terms of an overall pattern and not in terms of different values.

3.2 Identification of reference vectors

After creating the direction vectors, the algorithm divides them according to the (existing) labels of the respondent categories. This division is done in order to identify the *reference vector* of each category. The reference vector is the representative preference pattern specific to a respondent category. The procedure of reference vector identification is divided in two modules; (i) measurement of decision scores and, (ii) decision making procedure.

Measurement of Decision Scores. The reference vector of each category is identified one by one. The algorithm counts the frequencies of −1, 0, and 1 for each variable in the subset of direction vectors corresponding to a category. The "probabilistic scores" of these values are then calculated by the ratio of the frequency of a variable's direction value to the total number of direction vectors in the subset. If this proportion of a direction value (−1, 0, or 1) for a variable is equal or more than the "conformity constraint" $\alpha = 0.8$, then this direction value is considered as a "confirmed value" for that variable and the variable is considered as "determined". The conformity constraint has been set to 0.8 on the basis of experimental trials. If the probabilistic score of any specific direction value corresponding to a variable qualifies this constraint then it implies that this value is highly prevalent in the variable due to which the variable becomes nearly a constant. If no direction value for a variable qualifies this constraint then the variable is considered as "undetermined". In that case the frequency based proportion is calculated through a procedure defined as shown in the following Algorithm 1.

By applying Algorithm 1, probabilistic scores are calculated for all undetermined variables. Let us take our running example depicted in Figure 3. First, the method divides the direction vectors of the example data into three subsets based on the respondent categories. Now, the method tries to iden-

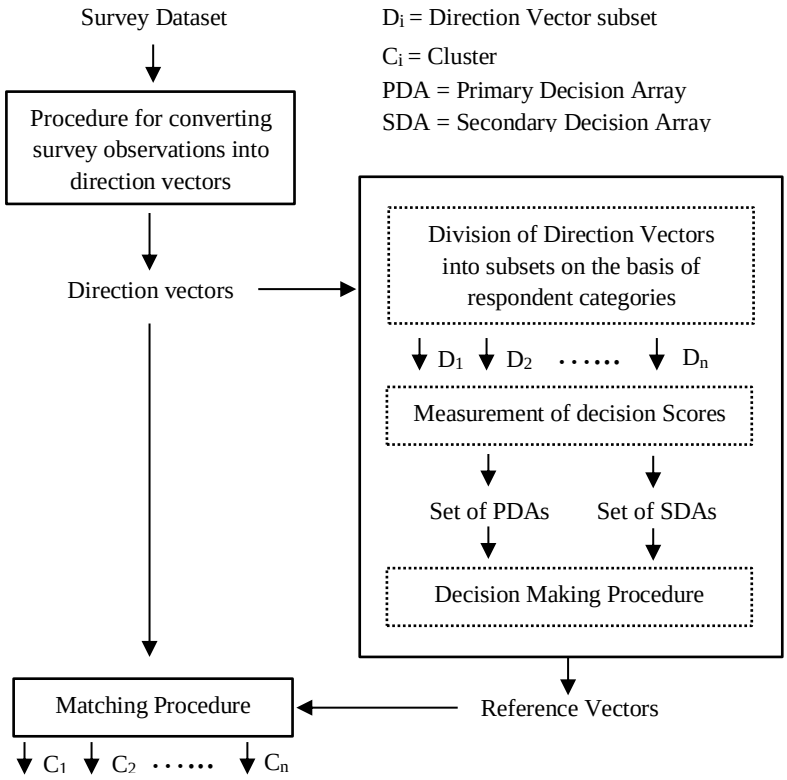
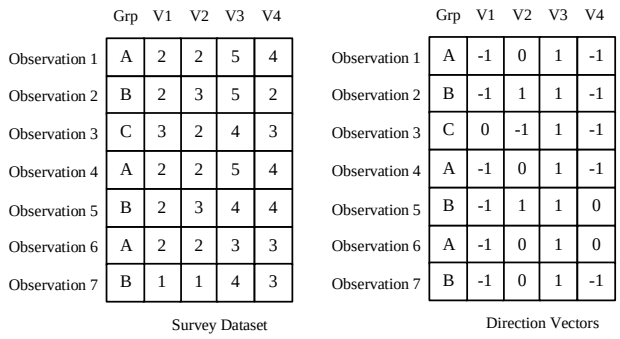


Figure 2. Architecture of proposed pattern clustering technique (PCSD3)



	Grp	V1	V2	V3	V4
Observation 1	A	2	2	5	4
Observation 2	B	2	3	5	2
Observation 3	C	3	2	4	3
Observation 4	A	2	2	5	4
Observation 5	B	2	3	4	4
Observation 6	A	2	2	3	3
Observation 7	B	1	1	4	3

Survey Dataset

	Grp	V1	V2	V3	V4
Observation 1	A	-1	0	1	-1
Observation 2	B	-1	1	1	-1
Observation 3	C	0	-1	1	-1
Observation 4	A	-1	0	1	-1
Observation 5	B	-1	1	1	0
Observation 6	A	-1	0	1	0
Observation 7	B	-1	0	1	-1

Direction Vectors

Figure 3. Direction vectors for the example data

tify the reference vector of each subset. We take the subset of direction vectors corresponding to category B for elaborating the mechanism. Figure 4 shows the steps involved to arrive from the direction vector data to the frequency counts and to the probabilistic score. We can see that proportion of the direction value -1 for V1 and for the direction value 1 for V3 are greater than α. Hence, -1 and 1 are confirmed values for V1 and V3 respectively. No value for V2 and V4 qualifies α, hence these are treated as undetermined. Now for V2 and V4 the probabilistic scores are calculated by Algorithm 1. For V2, the absolute frequency of 1 is largest (2) and for 0 it is second largest (1), hence according to the steps 7 and 8 of Algorithm 1, the least frequency (0) corresponding to value -1 is subtracted from the largest frequency ($R = 2 - 0 = 2$). The probability of R with respect to the second largest frequency is the required magnitude of the score ($S = \frac{2}{2+1} = 0.66$). The sign of S is positive as the value $+1$ has the largest frequency. Similarly, we find the score for the fourth variable, V4 (-0.66).

An intermediate vector—the "Primary Decision Array" (PDA)—is formed that contains confirmed values of the determined variables and probabilistic scores of the undetermined variables. The PDA for category B—$\{-1, 0.66, 1, -0.66\}$—is also shown in Figure 4. In total three PDAs are created by the method, one for each subset of the respondent category. The PDA corresponding to category A, $\{-1, 0, 1, -0.66\}$ is calculated analogously to the PDA of category B. Since, category C consists of a single observation the direction vector becomes its own reference vector.

Now, the method searches for relationships between the variables. This is done by observing the frequencies of direction values in all other variables with respect to each specific direction value of a chosen variable. For example, when the direction value of variable V1 of category B is -1, then the corresponding frequencies of values -1, 0, and 1 occurred in

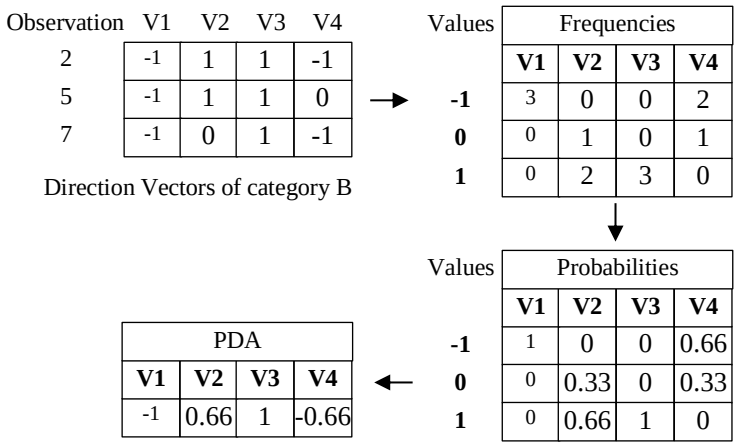


Figure 4. Primary Decision Array (PDA) for category B of the example data

V2, V3, V4 are counted. Similarly, frequency lists for direction values 0 and 1 of V1 are recorded. In this way three separate lists of frequencies are recorded for every variable. Each separate list is finally converged into a single vector by finding confirmed values and probabilistic scores with the help of the conformity constraint α, and Algorithm 1 (as we have done while creating PDA). All of these intermediate vectors corresponding to each variable in a category are jointly called "Secondary Decision Array" (SDA). Figure 5 shows the frequency lists for each variable for category B in the running example and its resulting SDA.

Decision Making Procedure. The PDA and SDA are utilized for identifying the reference vector. Determined variables in PDA are used to finalize undetermined variables. If no determined variable is found in a PDA, then constraint α is lowered by 0.05 and the PDA is reconstructed. If necessary, then α is lowered further by 0.05 till we get at least one single determined variable in PDA. Finalization of undetermined variables is done through adaptive decision making procedure which runs in several iterations and finalizes one undetermined variable in each iteration. Each time a variable is finalized, the PDA is reconstructed, and the finalized variable becomes available for the finalization process of other undetermined variable from next iteration. After finalizing all the variables, the resulting PDA is considered as the desired reference vector. The whole decision making procedure is shown in Algorithm 2.

Consider the PDA of category B of the running example (Figure 4). Here, V1 and V3 contain confirmed values. Now we look into the SDA for the values of V2 and V4 corresponding to −1 in V1 and 1 in V3 (first and third lists of SDA in Figure 5). We now have two values for each V2 and V4 that are 0.66 and 0.66 for V2, −0.66 and −0.66 for V4. Although the pair of values for each variable are the same in the very small example data, these values may differ based on the frequencies of direction values. The mean value from both of these variables is than calculated, which is 0.66 for V2 and −0.66 for V4. For deciding confirmed value of undetermined variables we have adopted a straightforward rule. We divided total value space (−1 to 1) in three parts. A score in the range −1 to −0.50 is mapped to the confirmed value −1. The ranges −0.49 to 0.49 represents 0, and finally range 0.5 to 1 corresponds to a 1. If a mean score lie between −1 and −0.50 than we check how close it is to −0.75. For score between −0.49 to 0 we check its closeness to −0.25. Same is followed for the positive values. The mean score of an undetermined variable showing the smallest distance with the threshold is chosen for finalization, and the score of it in the PDA is replaced with the corresponding confirmed value.

In our example, mean 0.66 of V2 and −0.66 for V4 are at equal distance to the predefined thresholds (0.75 for positive 1 and −0.75 for negative 1). Therefore, the method is free to choose between V2 and V4 for finalization and it chooses V2 according to its natural sequence. The score of V2 (0.66)

Algorithm 2 (Decision making with the help of PDA and SDA)

1: Identify determined and undetermined variables in PDA.
2: Fetch values of undetermined variables from SDA corresponding to determined values in PDA.
3: Calculate the mean from the set of SDA values corresponding to each undetermined variable.
4: Check, for which variable the mean value shows highest proximity with a confirmed value (-1, 0, 1).
5: Replace score with confirmed value in PDA for variable showing highest proximity.
6: Repeat from step 1 for modified PDA until all variables in PDA are confirmed.

V1		V2	V3	V4
-1	-1	0	0	2
	0	1	0	1
	1	2	3	0
0	-1	0	0	0
	0	0	0	0
	1	0	0	0
1	-1	0	0	0
	0	0	0	0
	1	0	0	0

V2		V1	V3	V4
-1	-1	0	0	0
	0	0	0	0
	1	0	0	0
0	-1	1	0	1
	0	0	0	0
	1	0	1	0
1	-1	2	0	1
	0	0	0	1
	1	0	2	0

V3		V1	V2	V4
-1	-1	0	0	0
	0	0	0	0
	1	0	0	0
0	-1	0	0	0
	0	0	0	0
	1	0	0	0
1	-1	3	0	2
	0	0	1	1
	1	0	2	0

V4		V1	V2	V3
-1	-1	2	0	0
	0	0	1	0
	1	0	1	2
0	-1	1	0	0
	0	0	0	0
	1	0	1	1
1	-1	0	0	0
	0	0	0	0
	1	0	0	0

Frequency lists of direction values corresponding to each particular value of variables

SDA

V1	V2	V3	V4
-1	0.66	1	-0.66
0	0	0	0
1	0	0	0

V2	V1	V3	V4
-1	0	0	0
0	-1	1	-1
1	-1	1	-0.50

V3	V1	V2	V4
-1	0	0	0
0	0	0	0
1	-1	0.66	-0.66

V4	V1	V2	V3
-1	-1	0.50	1
0	-1	1	1
1	0	0	0

Frequency based probabilistic scores corresponding to each particular value of variables

Figure 5. Secondary Decision Array (SDA) for category B of the example data

in PDA is replaced with 1 according to our mapping scheme. For finalizing V4, the whole procedure of Algorithm 2 is repeated, which now have three determined variables V1, V2, and V3 in the reconstructed PDA. The final iteration of the procedure sets the value of V4 to -1 and produces its final PDA for category B, which is $\{-1, 1, 1, -1\}$. This final PDA with all confirmed values is the desired reference vector of category B. The reference vectors of other two categories are created correspondingly. For the running example This leads to the following reference vectors:

$$\begin{aligned} \text{A:} &\quad \{-1, 0, 1, -1\} \\ \text{B:} &\quad \{-1, 1, 1, -1\} \\ \text{C:} &\quad \{0, -1, 1, -1\} \end{aligned}$$

These reference vectors are now forwarded to the matching procedure.

The main feature of adaptive decision making (Algorithm 2) is that the confidence of algorithm grows each time it completes its iteration. Since a new outcome participates in decision making of next iteration, the sample space of the algorithm grows per iteration. In our example, the first iteration of the algorithm has a set of two samples (V1, V3) to decide the inclination of undetermined variables (V2, V4). While in last iteration it has three samples (V1, V2, V3) to decide the value of a single variable (V4). This adaptability in decision making enhances the accuracy of the algorithm.

3.3 Matching

Before matching, each reference vector is given an index for the purpose of cluster naming. Since PCSD3 recognizes the respondent categories and their representative features (reference vectors) it names the clusters with the names of respondent categories. The observation is labeled with the name of the reference vector to which it's direction vectors shows the highest proximity. This name indicates the cluster to which the observation under consideration belongs. The dataset of the direction vectors, which was created during the first part, is matched to the reference vectors identified in the second part. Each direction vector is matched to all reference vectors, and the absolute distances to each reference vector are measured. Absolute distance is simply the sum of absolute differences $|x - y|$ in each dimension between two vectors. The absolute difference is chosen since our value space is constituted with three equidistant values (-1, 0, and 1). The smallest absolute distance of a direction vector from a particular reference vector shows highest similarity. Hence,

the index of the reference vector for which the smallest distance is found is used to label the original observation of that direction vector.

The matching procedure for the example data is shown in Figure 6. The explored reference vectors are labeled with the names of the respondent categories (A, B, C) they are representing. Each direction vector of the example data is then matched to the each of the three reference vectors, and the absolute distances between them are calculated. For instance, the direction vector of the first observation is matched with to all the three reference vectors (A, B, C) and its distances (0 with A, 1 with B, and 2 with C) are calculated. Now each of the seven observations is assigned to the label of a reference vector for which its direction vector shows the smallest absolute distance e.g., observation 1 is assigned with cluster label A as its distance with reference vector A is smallest (0). Following this way, all the seven observations of the example are separated in three clusters named as Cluster A, Cluster B, and Cluster C. The matching procedure clusters observations 1, 4, 6, and 7 in Cluster A, 2 and 5 in Cluster B, and observation 3 in Cluster C.

4 Stakeholder Theory & Applicability in HEIs

The stakeholder theory provides a foundation to recognize relevant entities inside the organization, and the logic for integrating their interests into decision making (Mitchell, Agle, & Wood, 1997). It states that an organization should concern about those, who have interests or stake in it while making strategic decisions (Argandoña, 1998; Freeman, 2010). Theory has gained enormous popularity as it is used in several domains of public interests i.e., policy science, education, health, and corporates (Brugha & Varvasovszky, 2000; Jensen, 2002).

Education sector uses stakeholder theory for various purposes. The assessment of the academic quality is one of its core applications (Vroeijenstijn, 2003). In the education sector various stakeholders have different perceptions of the academic quality (A. Burrows & Harvey, 1992) so that the academic quality is seen as the difference between the stakeholder's expectations and the actual perceived performance of an institution (Athiyaman, 1997; Bourner, 1998; O'Neill & Palmer, 2004). Recent developments regarding exploration of quality parameters of HEIs also reflects significant use of stakeholder theory (Akareem & Hossain, 2012; Senthilkumar & Arulraj, 2011; Vnouckova, Urbancová, & Smolová, 2017). Overall, stakeholder theory is well established in research on education.

Literature related to educational quality has encountered several kinds of stakeholder groupings e.g., grouping based on relationships, based on salience over HEIs etc., (J. Burrows, 1999; Jongbloed, Enders, & Salerno, 2008; Lyytinen, Kohtamäki, Kivistö, Pekkola, & Hölttä, 2017). However, most of the studies regarding quality of HEIs consider students, faculty, parents, and administrators as key academic stakeholders (Aydin, 2013; Iacovidou, Gibbs, & Zopiatis, 2009; Telli, 2013). Since stakeholders groups are defined on the basis of their divergent interests, a grouping scheme is valid if a decent majority of population in each group shares similar interests based on which it is made. Since quality parameters of HEIs are defined according to the preferences of academic stakeholders, any clustering based on preference patterns should divide the data in its natural stakeholder groups. This implies that clustering should divide data in such a way that each distinct cluster represents a specific stakeholder category. With respect to this hypothesis we lay down the following research question:

RQ: Is there a fair proportion of the population in each defined stakeholder group that follows similar preferences with respect to quality parameters of HEIs?

5 Data

The data was collected for a study to explore the relevance of various quality parameters of HEIs (see replication materials). Eleven items for such quality parameters were explored in the study. Detailed information regarding these parameters and their basic statistics are given in Section 6. Six of these items were selected after exhaustive scrutiny of the five most popular international and national institutional rankings. Five additional items were created by means of focus group and personal interviews of students, faculty, parents, administrators, and professionals. An survey was fielded to measure the perception of a large sample of stakeholders on these parameters. The survey used Likert items with four levels to evaluate the relevance of the parameters.

The study was fielded in the National Capital Region (NCR) of India since NCR has representative premier institutions, and the population in these institutions is assumed to represent whole India. The survey invited respondents from Sciences, Social-Sciences, Medical, Technology, and Humanities domains of twelve premier HEIs of the country. Data of seven respondent categories were collected, namely undergraduate students, graduate students, graduate researchers, faculty, parents, administrators, and professionals. The population in each category except faculty and administrator was assumed to be infinite. The overall population of faculty in the chosen institutions was 5727, whereas no official data was found for the population of administrators (NIRF, 2018). In total 2620 respondents could be used for the study, with 438 undergraduates, 463 graduate students, 447 graduate researchers, 389 professionals, 395 parents, 401 faculty, and 87 administrators.

The actual population of the administrator category in the selected institutions is not known and the number of responses from respondents of this category is much smaller than that of the other categories. The reason for the low number is that most administrators have dual roles: academic and

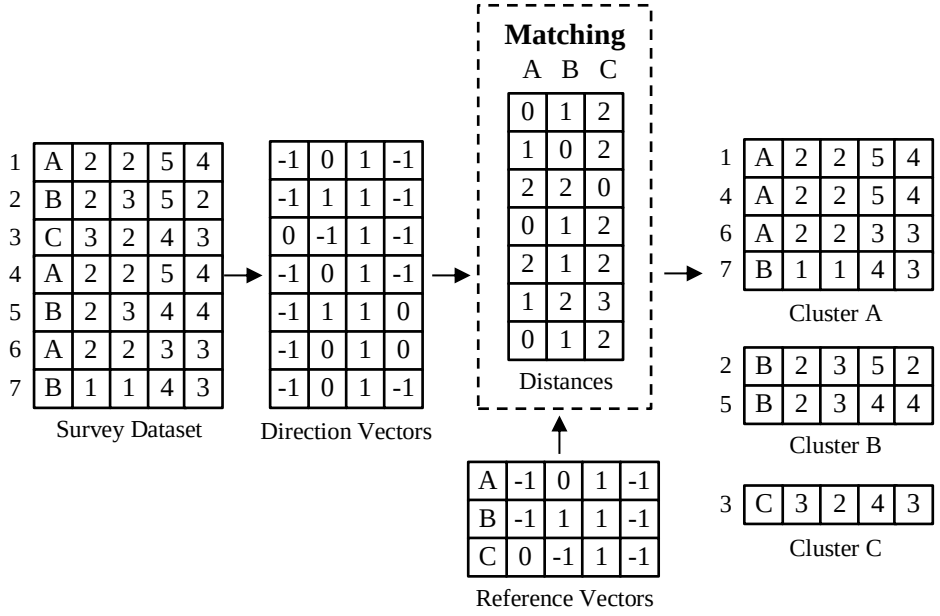


Figure 6. Matching procedure and cluster formation of the example data

Table 1
Means of 11 items for quality parameters by respondent category

		Respondent category						
Quality Parameters	Overall	UG	GS	GR	FAC	PRO	PAR	AD
Teaching	3.2	3.4	3.0	3.3	3.5	3.1	3.2	2.9
Graduate Outcomes	3.0	3.3	3.0	2.3	2.5	3.3	3.4	3.0
Academic Flexibility	3.0	3.0	3.3	3.1	3.0	3.0	2.6	2.2
Transparency & Accountability	3.0	2.8	2.9	3.2	3.1	2.9	3.0	3.1
Infrastructure & Resources	3.0	3.2	3.0	2.9	2.7	2.8	3.1	3.0
Research	3.0	2.0	2.7	3.4	3.4	3.2	3.0	3.0
Student Support Services	2.9	3.2	3.1	2.9	2.4	2.7	3.1	2.5
International Outlook	2.8	2.4	2.8	3.1	3.0	2.8	3.0	2.4
Fee & Financial Assistance	2.7	3.1	2.7	3.2	2.0	2.0	3.3	1.8
Academic Autonomy	2.5	2.0	2.3	2.7	3.2	2.1	2.6	3.4
Inclusivity	2.5	2.5	2.3	2.8	2.4	2.1	2.7	2.5

administrative. The survey counted these persons as faculty so that administrators are only full-time non-faculty administrators. Except for the administrator category, the remaining data is quite balanced and satisfies some minimum sampling requirements.

For the cluster analysis presented in this paper, we excluded the administrator category. Results including the administrators are shown in the appendix.

6 Data Analysis & Clustering Results

The following subsection reports descriptive statistics for the items on of the quality parameters. The results of the cluster analysis are presented thereafter. We thereby compare the results of PCSD3 and *K*-means against the expectations of the stakeholder theory. To ease the representation, the respondent categories used in the analysis are abbreviated as follows:

UG	Undergraduates
GS	Graduate Students
GR	Graduate Researchers
FAC	Faculty
PAR	Parents as PAR
PRO	Professionals
AD	Administrators

6.1 Statistics of Dataset Used

Table 1 shows the overall and category-wise survey means of the 11 items for the quality parameters. The higher the mean the higher is the perceived relevance of the corresponding quality parameter. Since the answer categories of all the items range from one to four, we consider means above the 2.5 as relevant. In this sense, all quality parameters are perceived as relevant by the academic stakeholders.

There is a huge variation in the means between respondent categories. This implies that priorities of different stakeholder categories are quite different.

The impact of aggregation can also be observed from the results of Table 1 as five of the eleven parameters in overall column shares same mean value (3.0), although there is substantial variability regarding these parameters between the respondent categories. This indicates that the treatment of small ordinal values on the basis of magnitudes (aggregation) suppresses patterns which may be important for behavioral studies.

6.2 Cluster Analysis

The following results of the cluster analysis excluded administrators and is thus based on only six of the seven respondent categories.

Clustering by PCSD3. PCSD3 clusters are labeled with the name of respondent categories. As outlined in section 3, PCSD3 utilizes representative choice patterns of categories and labels clusters according to the names of these categories. This is one of the striking features of PCSD3 that makes the recognition of clusters easier and makes results easier to interpret. An added value of labeling the clusters with existing groups in data is the accessibility ease of performance evaluations. Subsequent paragraphs show how such labeling helps in calculating various evaluating parameters.

Table 2 shows the classification of respondents into the clusters by respondent category. Reading the table horizontally (row-wise) informs about the frequencies of respondent from different categories in a cluster. For an instance, row 4 of Table 2 shows that the cluster FAC contains 22 undergraduates, 41 graduate students, 74 researchers, 279 faculty, 56 professionals, and 18 parents. Overall, 490 respondents belong to cluster FAC. Reading the table vertically (column-wise) informs about the distribution of respondent categories into the PCSD3 clusters. The second column, for example, shows the clusters in which undergraduates end up.

We can gather from the horizontal interpretation of Table 2 that the proportion of faculty in cluster FAC is quite high (57%). This means that the preferences of the cluster FAC correspond to a large extent to the preferences of faculty; we say the cluster is "specific". Similarly, the proportions of both, professionals in cluster PRO, and graduate researchers in cluster GR is 52%. Such high proportions suggest that these clusters are specific, too. Proportions of undergraduates in cluster UG and parents in cluster PAR are 44% and 41% respectively, so that the relative majority of respondents in these clusters also stem from their representative population. The lowest value for this kind of specificity of a cluster is found for cluster GS, in which the representative population has a proportion of around 34%. While this is comparatively low, it is a fair majority and still capable to clearly distinguish the cluster.

Overall, clustering done through PCSD3 clarifies that each cluster labeled with a particular respondent category contains a relative majority of responses from that category. This phenomenon validates the applicability of used stakeholder grouping and proves our hypothesis. It implies that the academic communities have their specific preferences and a fair proportion of their respective members follow similar patterns.

The vertical interpretation of Table 2) also gives relevant information. It represents the alignment of member's choices to the representative preference pattern of their own community. There is a quite high proportion (70%) of faculty members in the cluster FAC. This suggests that a huge majority of faculty members follow the same pattern of choices. The reason is clear, faculty members are well adapted to academia, they have clear vision about the requirements and thus strongly aligned opinions. The distribution of professionals, parents and undergraduates also show large proportions ($\approx 50\%$) of their population in their representative cluster. In that sense they are also aligned in their opinions, although to a lesser degree. In contrast, graduate students and researchers possess divergent behaviors as their distribution in their representative clusters are somewhat low (around 30–35%). While such proportions allow to distinguish their choices from the other respondent categories, the fair proportions in the other clusters suggests that they are less aligned to their opinions. This is likely due to their divergent aspirations.

Table 2 can be also taken as the *confusion matrix* for the proposed method, where category labels in the header row represent actual respondent categories in data and rows indicate the categories predicted by the method. One can thus directly calculate measures such as true positives, true negatives, etc. Take the respondent category faculty in Table 2 for instance. The true positives for this case are found at the intersection of the actual and predicted respondent category for FAC (279). The false positives are the sum of all values found in predicted class FAC, excluding the true positives (i.e. 22 + 41 + 74 + 56 + 18 = 211). False negatives are the sum of all values found in actual class FAC excluding the true positives (i.e. 18 + 34 + 26 + 12 + 32 = 122). The true negatives are the sum of all values in matrix excluding the values in the row of predicted class FAC (1921, i.e. the sum of all values excluding the values in fourth row). Since,

Table 2
Number of respondents in PCSD3 clusters by respondent category (Confusion matrix of PCSD3)

	Respondent category						
Clusters	UG	GS	GR	FAC	PRO	PAR	Total
UG	215	99	35	18	42	83	492
GS	59	141	81	34	44	62	421
GR	13	72	159	26	29	10	309
FAC	22	41	74	279	56	18	490
PRO	55	68	5	12	191	36	367
PAR	74	42	93	32	27	186	454

Table 3
Performance measures for PCSD3 solution

Categories	TP	TN	FP	FN	Precision	Recall	Accuracy
UG	215	1818	277	223	0.44	0.49	0.80
GS	141	1790	280	322	0.33	0.30	0.76
GR	159	1936	150	288	0.51	0.36	0.83
FAC	279	1921	211	122	0.57	0.70	0.87
PRO	191	1968	176	198	0.52	0.49	0.85
PAR	186	1870	268	209	0.41	0.47	0.81
Mean	195	1884	227	227	0.46	0.47	0.82

Table 4
Number of respondents in K-means clusters by respondent category (Confusion matrix of K-means)

	Respondent category						
Clusters	UG	GS	GR	FAC	PRO	PAR	Total
C1	238	45	17	8	20	99	427
C2	39	80	120	36	76	5	356
C3	25	94	144	56	63	98	480
C4	20	17	41	227	34	26	365
C5	59	135	10	20	175	25	424
C6	58	92	115	54	21	142	481

Table 5
Performance measures for k-means solution

Categories	TP	TN	FP	FN	Precision	Recall	Accuracy
C1(UG)	238	1906	189	200	0.56	0.52	0.83
C2(GS)	80	1794	276	383	0.22	0.17	0.73
C3(GR)	144	1750	336	303	0.30	0.32	0.75
C4(FAC)	227	1994	138	174	0.62	0.56	0.87
C5(PRO)	175	1895	249	214	0.41	0.44	0.82
C6(PAR)	142	1798	340	253	0.29	0.35	0.80
Mean	168	1856	255	258	0.40	0.40	0.80

PCSD3 labels the clusters with the name of the respondent categories, performance measures such as precision, recall, and accuracy of the method can be easily calculated. This is not possible with other clustering methods as they divide data in unsupervised fashion so that the resulting clusters do not have direct links to existing categories of data.

Table 3 shows the results of some performance measures calculated from the confusion matrix of Table 2. The results are compared to the corresponding results of the K-means solution in the following subsection.

Clustering by K-means. Table 4 shows the results of K-means clustering for the same data. Equivalently to Table 2, it shows the classification of respondents into the clusters by respondent category. The clusters are labeled by their respective cluster indices, i.e., C1, C2, ..., C6. The cluster indexing followed the aiming to create a confusion matrix. Since, K-means clusters data in an unsupervised manner there is no predefined way to label the clusters and the resulting clusters are thus free of any pre-existing categorization. Thus, the clusters have to be evaluated on the basis of ground truth of the problem. The ground truth—here in our case—is stakeholder theory, according to which natural clustering should segregate the data on behalf of respondent categories having divergent preferences. This is also the reason due to which we choose $K = 6$, that is, one for each respondent category. Moreover, the comparison of PCSD3 with K-means is only possible for $K = 6$. Therefore, we interpret the K-means clusters using the proportions of respondent categories in the clusters.

According to the major proportion of categories, K-means cluster C1 corresponds to undergraduates, and C4 corresponds to faculty. The remaining clusters cannot be linked to a specific respondent category with sufficient confidence as some of these clusters contain large numbers of multiple respondent categories. For example, cluster C5 contains high numbers form both, professionals and graduate students. Moreover, a large proportion of graduate students and graduate researchers are distributed among multiple clusters. For instance, researchers are distributed to clusters C3 and C2 in almost equal proportions of 34%, and 30% respectively. Due to these reasons, K-means clusters cannot be easily interpreted for the data at hand. In the context of our own study, the K-means cluster solution does not produce relevant information.

Since the K-means clusters cannot be clearly linked to the respondent categories, the evaluation of the cluster solution requires some contradictable settings. We linked most of the respondent categories to the clusters in which their highest proportion is contained, leaving one exception (graduate students). C1 is linked to undergraduates, C2 to graduate students, C3 with graduate researchers, C4 with faculty, C5 with professionals, and C6 to parents. Using this labeling scheme, Table 4 can be used as a confusion matrix. As in the previous subsection, Table 5 shows the results of performance measures calculated from this confusion matrix.

The comparison of PCSD3 and K-means regarding the accumulation of respondent categories (Tables 2 and 4) shows that PCSD3 outperforms K-means in segregating respondent categories. One can easily distinguish the clusters from PCSD3 by the distribution of respondent categories in the clusters, while this is not true for K-means. The distribution of clusters by respondent categories (vertical interpretation) also suggests that respondents tent to end up their representative PCSD3 cluster. This is also not true for the K-means solution.

The comparison of PCSD3 and K-means using the various performance parameters portrays the same story from a different perspective. Except for one or two individual cases such as undergraduates, PCSD3 outperforms the K-means solution. The overall comparison on the basis of mean value of each performance parameter (given in the bottom row of Tables 3 and 5) suggests that PCSD3 outperforms K-means in every possible ways.

The reason for the better performance of PCSD3 is that it's mechanism is specifically designed according to the features of survey data containing ordinal values. PCSD3 utilizes directional difference of the whole observation instead of magnitude difference; hence it avoids the suppression of delicate patterns which are quite important for depicting the differences in preferences inside survey data. Overall, the results suggest that PCSD3 is suitable for survey data, and that its clusters are more interpretable.

6.3 Discussion

The results of the proposed PCSD3 method and its comparison with K-means method suggest that PCSD3 works suitably well for survey data containing ordinal values. Since it works on pattern based similarity it avoids the limitations of value based (magnitude dependent) measurement of small ordinal values. Further, it does not require arbitrary manual settings of clustering parameters such as the number of clusters. Instead, it takes advantage of theoretically justifiable groupings in the data. The clusters made by PCSD3 are more interpretable due to the automated linkage of the clusters to those groupings. As an added value, PCSD3 can be used as pattern filter for applications where previously known patterns (behaviour or preferences) are sought in data. We claim that PCSD3 is especially useful for organizational and behavioral research. It can be used to analyze the appropriateness of various kinds of groupings and to study general tendency over long lists of items for designated groups of persons.

A limitation of PCSD3 is that its results vary with change in the sequence of variables. The reason behind such order dependence is the use of directions. Change in variable order produce change in directions therefore the results are bound to vary. Best results therefore, can be achieved by arranging

the variables in a way that maximize variability among them. The applicability of PCSD3 is limited to survey data having ordinal values, it cannot handle nominal and mixed data. Further, PCSD3 is limited to the applications that seeks pattern-wise recognition of trends with respect to existing categories in data.

Despite those limitations, the method opens up the direction of research for developing more sophisticated pattern clustering methods through coupling its newly developed mechanisms (frequency based probabilistic scores and adaptive decision making) with machine learning.

7 Conclusion

This paper proposed a pattern clustering method (PCSD3) specially designed for survey data containing ordinal values. The method works on the basis of directional differences and uses frequency based probabilistic scoring to avoid the limitations of magnitude based treatment of small ordinal values. It does not require manual setting of clustering parameters but uses theoretically justifiable groupings in the data for that purpose.

We applied PCSD3 on real survey data and compared it with a *K*-means clustering method. PCSD3 divided this dataset into clusters that differ strongly between stakeholder groups. This phenomenon illustrates that PCSD3 is capable to segregate the data into pre-defined grouping whereas this is not true for *K*-means. This implies that the proposed clustering method works suitably well for quantitative survey data. Therefore, the results empirically validates earlier studies that advocated the use of pattern based measures in organizational and behavioral studies.

Acknowledgment

We thank the reviewers for their insightful comments and suggestions by which the understanding and readability of this manuscript is improved.

The data and material used in this work is available in the replication materials and through a *GitHub* repository.[1]

There is no known conflict of interests associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

References

Akareem, H. S., & Hossain, S. S. (2012). Perception of education quality in private universities of Bangladesh: A study from students' perspective. *Journal of Marketing for Higher Education*, *22*(1), 11–33.

Argandoña, A. (1998). The stakeholder theory and the common good. *Journal of Business Ethics*, *17*(9-10), 1093–1102.

Athiyaman, A. (1997). Linking student satisfaction and service quality perceptions: The case of university education. *European Journal of Marketing*.

Aydin, H. (2013). Four stakeholder's perception on educational effectiveness of Nigerian Turkish International colleges: A qualitative case study. *SAGE Open*, *3*(2).

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, *43*(3), 800.

Bourner, T. (1998). More knowledge, new knowledge: The impact on education and training. *Education + Training*, *40*(1).

Brugha, R., & Varvasovszky, Z. (2000). Stakeholder analysis: A review. *Health Policy & Planning*, *15*(3), 239–246.

Burrows, A., & Harvey, L. (1992). Defining quality in higher education: The stakeholder approach. In *Proc. AETT Conf. on Quality in Education* (pp. 6–8).

Burrows, J. (1999). Going beyond labels: A framework for profiling institutional stakeholders. *Contemporary Education*, *70*(4), 5.

Cheng, Y., & Church, G. M. (2000). Biclustering of expression data. In *Proc. ISMB* (Vol. 8, pp. 93–103).

Church, A. H., & Waclawski, J. (2001). *Designing and Using Organizational Surveys: A Seven-Step Process*. John Wiley & Sons.

Estivill-Castro, V. (2002). Why so many clustering algorithms: A position paper. *ACM SIGKDD Explorations Newsletter*, *4*(1), 65–75.

Freeman, R. E. (2010). *Strategic Management: A Stakeholder Approach.* Cambridge University Press.

Grice, J. W. (2014). Observation oriented modeling: Preparing students for research in the 21st century. *Comprehensive Psychology*, *3*, 05–08.

Grice, J. W. (2015). From means and variances to persons and patterns. *Frontiers in Psychology*, *6*, 1007.

Guha, S., Rastogi, R., & Shim, K. (2000). ROCK: A robust clustering algorithm for categorical attributes. *Information Systems*, *25*(5), 345–366.

Hill, C. A., Biemer, P., Buskirk, T., Callegaro, M., Cazar, A. L. C., Eck, A., ... Lyberg, L., et al. (2019). Exploring new statistical frontiers at the intersection of survey science and big data: Convergence at "BigSurv18". *Survey Research Methods*, *13*(1), 123–134.

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, *100*(3), 828.

Iacovidou, M., Gibbs, P., & Zopiatis, A. (2009). An exploratory use of the stakeholder approach to defining and measuring quality: The case of a Cypriot higher

[1] https://github.com/RoopamSadh/PCSD3

education institution. *Quality in Higher Education*, *15*(2), 147–165.

Jensen, M. C. (2002). Value maximization, stakeholder theory, and the corporate objective function. *Business Ethics Quarterly*, *12*(2), 235–256.

Jiang, D., Tang, C., & Zhang, A. (2004). Cluster analysis for gene expression data: A survey. *IEEE Trans. Knowledge & Data Engineering*, *16*(11), 1370–1386.

Jongbloed, B., Enders, J., & Salerno, C. (2008). Higher education and its communities: Interconnections, interdependencies and a research agenda. *Higher Education*, *56*(3), 303–324.

Kern, C., Klausch, T., & Kreuter, F. (2019). Tree-based machine learning methods for survey research. *Survey Research Methods*, *13*(1), 73–93.

Kriegel, H.-P., Kröger, P., & Zimek, A. (2009). Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. Knowledge Discovery from Data*, *3*(1), 1–58.

Lee, J. W., Jones, P. S., Mineyama, Y., & Zhang, X. E. (2002). Cultural differences in responses to a Likert scale. *Research in Nursing & Health*, *25*(4), 295–306.

Liu, J., & Wang, W. (2003). Op-cluster: Clustering by tendency in high dimensional space. In *Proc. 3rd IEEE Int. Conf. Data Mining* (pp. 187–194). IEEE.

Lyytinen, A., Kohtamäki, V., Kivistö, J., Pekkola, E., & Hölttä, S. (2017). Scenarios of quality assurance of stakeholder relationships in Finnish higher education institutions. *Quality in HigHer Education*, *23*(1), 35–49.

Manicas, P. T. (2006). *A Realist Philosophy of Social Science: Explanation and Understanding.* Cambridge University Press.

Mitchell, R. K., Agle, B. R., & Wood, D. J. (1997). Toward a theory of stakeholder identification and salience: Defining the principle of who and what really counts. *Academy of Management Review*, *22*(4), 853–886.

NIRF. (2018). Ranking-2018. https://www.nirfindia.org/Home.

O'Neill, M. A., & Palmer, A. (2004). Importance-performance analysis: A useful tool for directing continuous quality improvement in higher education. *Quality Assurance in Education*, *12*(1).

Pei, J., Zhang, X., Cho, M., Wang, H., & Yu, P. S. (2003). Maple: A fast algorithm for maximal pattern-based clustering. In *Third ieee international conference on data mining* (pp. 259–266). IEEE.

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. d. F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PloS One*, *14*(1).

Sadh, R., & Kumar, R. (2020). Clustering of quantitative survey data: A subsystem of EDM framework. In *Advances in Intelligent Systems and Computing*, Springer.

Senthilkumar, N., & Arulraj, A. (2011). SQM-HEI–determination of service quality measurement of higher education in india. *Journal of Modelling in Management*, *6*(1), 60–78.

Tan, P.-N., Steinbach, M., Kumar, V., et al. (2006). Cluster analysis: Basic concepts and algorithms. *Introduction to Data Mining*, *8*, 487–568.

Telli, G. (2013). How should quality of education be redefined for education achievements in Tanzania? what are stakeholders' opinions?. *Journal Int. Education & Leadership*, *3*(1), n1.

Valsiner, J., Molenaar, P. C., Lyra, M. C., & Chaudhary, N. (2009). *Dynamic Process Methodology in the Social and Developmental Sciences.* Springer.

Van Mechelen, I., Bock, H.-H., & De Boeck, P. (2004). Two-mode clustering methods: Astructuredoverview. *Statistical methods in medical research*, *13*(5), 363–394.

Vandervalk, A., Louch, H., Guerre, J., & Margiotta, R. (2014). *Incorporating reliability performance measures into the transportation planning and programming processes*.

Vnouckova, L., Urbancová, H., & Smolová, H. (2017). Factors describing students' perception on education quality standards. *Journal on Efficiency and Responsibility in Education & Science*, *10*(4), 109–115.

Vroeijenstijn, A. (2003). Towards a quality model for higher education. *Journal of Philippine Higher Education Quality Assurance*, *1*(1), 78–94.

Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S., et al. (2001). Constrained *k*-means clustering with background knowledge. In *Proc. ICML* (Vol. 1, pp. 577–584).

Wang, H., Chu, F., Fan, W., Yu, P. S., & Pei, J. (2004). A fast algorithm for subspace clustering by pattern similarity. In *Proc. 16th Int. Conf. Scientific and Statistical Database Management* (pp. 51–60). IEEE.

Wang, H., & Pei, J. (2008). Clustering by pattern similarity. *Journal of Computer Science & Technology*, *23*(4), 481–496.

Wang, H., Wang, W., Yang, J., & Yu, P. S. (2002). Clustering by pattern similarity in large data sets. In *Proc. ACM SIGMOD Int. Conf. Management of Data* (pp. 394–405).

Ward, M., & Meade, A. W. (2018). Applying social psychology to prevent careless responding during online surveys. *Applied Psychology*, *67*(2), 231–263.

Xing, E. P., Jordan, M. I., Russell, S. J., & Ng, A. Y. (2003). Distance metric learning with application to clustering with side-information. In *Proc. Conf. Advances in Neural Information Processing Systems* (pp. 521–528).

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*(2), 165–193.

Yang, J., Wang, W., Wang, H., & Yu, P. (2002). /spl delta/-clusters: Capturing subspace correlation in a large data set. In *Proceedings 18th international conference on data engineering* (pp. 517–528). IEEE.

Appendix A
Additional results

We are providing additional results of clustering over used dataset (including the responses of administrator category) in this section. The purpose is to present the general behavior of administrators and to evaluate the performance of proposed PCSD3 method in a different scenario.

It can be deduced from Table B1 that the PCSD3 method properly outlines each category in their representative clusters except for administrators. While exact conclusions cannot be made due to the small number of observations, the result suggests that administrators have quite divergent interests as their responses are distributed in several categories with fairly equal proportions. This implies that priorities of administrators regarding qualities of HEIs are quite idosyncratic, and do not follow any specific pattern.

The results of K-means ($K = 7$) shown in Table B2 are quite different. According to the proportions, undergraduates and professionals can be linked with clusters C3 and C7 respectively. Remaining categories cannot be linked with clusters due to their equal distribution in multiple clusters and/or the accumulation of multiple majorities in one cluster. One interesting phenomenon however can be seen in case of administrators. K-means cluster C1 contains majority of both, administrators and faculty, while the distribution of administrators is scattered into multiple categories according to the results of PCSD3. In this case the results of K-means are grossly misleading due to its sensitivity towards high magnitude differences.

In case of K-means, overall similarity in pattern is compromised if any specific dimension shows high differences in magnitudes. In other words, if an observation is highly dissimilar with a mean centroid in any specific dimension, then its overall similarity with respect to rest of the dimensions is compromised. For elaborating this phenomenon we depict mean value patterns of administrators, faculty, and professionals by standard error plots in Figure C1. We can see in plots that for Academic Autonomy (AA) the trends of both faculty and administrators are highly similar. While for other variables mean pattern of administrators is fairly similar with professionals. Due to strong similarity with respect to variable AA, K-means groups administrator's responses along with faculty (C1 of Table B2). In contrast, PCSD3 uses directional values instead of magnitudes, which neutralize the effect of isolated large differences. Since, it is obvious that strongly similar opinion regarding single dimension does not conclude overall similarity hence, results of PCSD3 are more reliable as it considers overall similarity in the pattern.

Appendix B
Tables

Table B1
Number of respondents in PCSD3 clusters by respondent category (Confusion matrix of PCSD3)

	Respondent category							
Clusters	UG	GS	GR	FAC	PRO	PAR	AD	Total
UG	248	85	15	10	31	78	4	471
GS	46	108	62	20	38	49	0	323
GR	10	72	159	26	27	10	6	310
FAC	18	37	58	251	31	18	15	428
PRO	40	64	1	8	164	31	20	328
PAR	57	37	87	24	27	181	23	436
AD	19	60	65	62	71	28	19	324

Table B2
Number of respondents in K-means clusters by respondent category (Confusion matrix of K-means)

	Respondent category							
Clusters	UG	GS	GR	FAC	PRO	PAR	AD	Total
C1	15	18	44	199	13	28	64	381
C2	22	45	73	26	65	5	1	237
C3	261	37	0	6	16	72	0	392
C4	5	69	131	10	13	64	0	292
C5	72	107	57	20	24	117	17	414
C6	27	71	136	94	60	88	5	481
C7	36	116	6	46	198	21	0	423

Appendix C
Figures

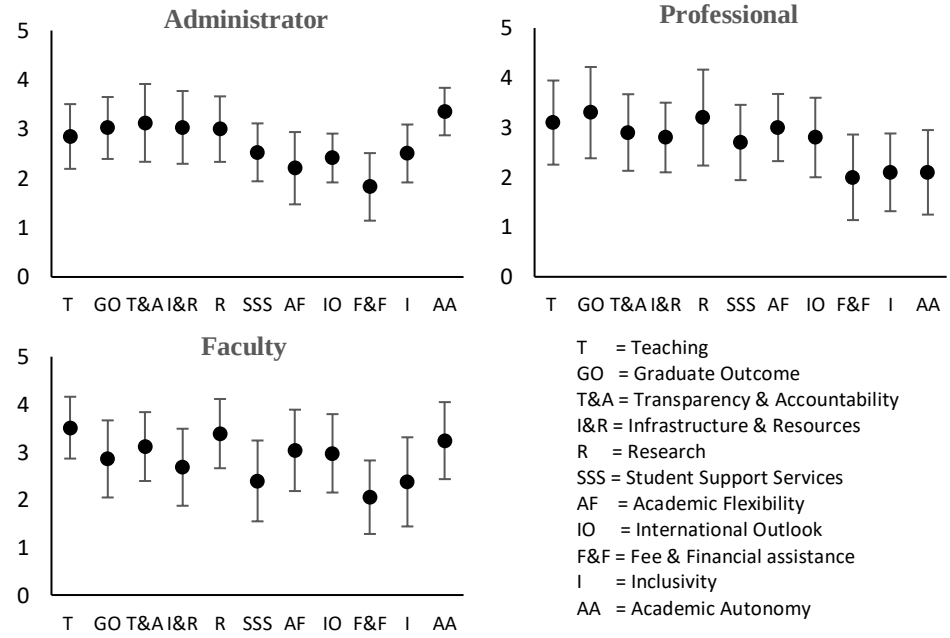


Figure C1. Pattern of mean values for different respondent categories.