Clustering of Quantitative Survey Data based on Marking Patterns

ROOPAM SADH[1]
RAJEEV KUMAR[2]

School of Computer & Systems Sciences
Jawaharlal Nehru University
New-Delhi - 110067, India
[1]roopam.sadh@gmail.com
[2]rajeevkumar.cse@gmail.com

Abstract. Clustering of quantitative survey data is done in-order to identify the divergent and dominant behaviors of the respondents. It is intended to explore the general tendencies of the respondent groups. Popular clustering methods working on value based similarity are inappropriate for survey data due to its distinct properties. Since marking patterns in survey data represents respondent's behavior hence separating the responses on the basis of marking patterns is an effective approach to identify the dominant behaviors. Thus, in this paper, we propose a specialized clustering method for quantitative survey data that combines the features of both, value based as well as pattern based approaches in order to obtain meaningful results. The proposed method does not require presetting of the clustering parameters while it makes use of group labels for selecting features and guiding the centroids at positions, which best describe divergent marking habits. We apply the proposed method over an educational survey dataset and compare its results with K-means clustering with respect to the benchmark stakeholder theory. Comparison results show that the proposed method is more appropriate for quantitative survey data.

Keywords: Cluster analysis, Guided clustering, Quantitative data, Stakeholder theory, Survey data.

(Received Sep. 10, 2020 / Accepted Nov. 15, 2020)

1 Introduction

The usefulness of a particular clustering method lies in the eyes of beholders [12]. Further, the suitability of a clustering method depends upon the properties of data [30, 43, 29]. For example, partition based clustering such as K-means are not suitable for non-convex data, whereas distribution based clustering such as DBSCAN are not suitable for sparse datasets with varying density [20]. Most of such clustering methods that work on value based similarity, use aggregate statistics, like mean, variance etc, which are found to be inappropriate for behavioural studies [14]. The survey data have some distinct properties, such as, fixed small range of ordinal values and associated information in the form of group labels [35]. Due to such properties, most of the value based clustering methods are not suitable for survey data. The desirable features of the value based clustering are their simplicity and computing efficiency [42], therefore, the pattern based clustering, on the other hand, is suitable under these circumstances as patterns in survey data reflects marking habits or behavior of the respondents [7]. The existing pattern based clustering methods are designed for specific applications and type of data, e.g., micro array analysis of gene expression data [18, 21, 41, 40] etc. Therefore these existing pattern clustering methods are not suitable for the quantitative survey data. Due to these reasons quantitative survey data [16, 33] requires a specialized clustering method for its proper segregation.

Therefore, we, in this paper, propose a clustering method for quantitative survey data that includes features of both: the value based as well as pattern based clustering methods. Further, it does not require pre-

setting of clustering parameters. The proposed method treats each survey observation as a single pattern instead of different variable values. It utilizes respondent group labels for deciding the number of clusters and feature selection. First, it divides explored patterns on the basis of their occurrence in each respondent group. Then, it finds mean marking vector of each subset of the patterns. These mean vectors are finally utilized as centroids for cluster the dataset. This way, the proposed method guides the mean centroids at positions in dataspace that best describe dominant marking behaviours of differing groups. The proposed method, thus, combines simplicity of value based clustering with effectiveness of pattern based clustering as it utilizes pattern based approach for feature selection and mean based centroids for clustering purpose.

We also compare the proposed method with K-means method on behalf of benchmark stakeholder theory [13, 25]. We apply the proposed clustering and K-means clustering over an original survey dataset which contains responses of different academic stakeholders with respect to the various quality parameters of higher educational institutions (HEIs). In this dataset, the stakeholder categories are defined on the basis of stakeholder's divergent interests (e.g., according to their roles) [5, 6, 24]. Therefore, a straightforward reasoning suggests that a natural clustering of survey responses should satisfy the stakeholders' grouping. Results of the our proposed clustering method over the dataset satisfy the stakeholder theory quite well, whereas the same is not true in case of K-means method. This indicates that the clusters made by the proposed method are comparatively more meaningful in context of quantitative survey data.

The rest of the paper is organized as follows. Section 2 gives a brief description of used dataset. The proposed clustering method is explained in Section 3. Analysis of the clustering results is included in Section 4. Salient features of the proposed clustering along with the future work are described in Section 5. Section 6, finally, concludes the findings of this work.

2 Used Dataset

The dataset, used in this work, was collected during a study that was intended to explore the quality parameters of HEIs and their relative importance [35]. Eleven quality parameters were discovered in the study. Six of these parameters were explored by rigorous scrutiny of five most popular international and national institutional rankings. These are: QS World University Rankings, Times Higher Education University Rankings, Academic Ranking of World Universities, The Complete University Guide (UK) and National Institutional Ranking Framework (India) [27, 38, 36, 9, 26]. Besides these six factors, five additional parameters were explored by conducting focus group and personal interviews of students, parents, administrators, faculty, and professionals. After discovering the parameters, perceptions of a large sample of academic stakeholders were explored regarding these parameters through an extensive online survey. These parameters are: Teaching (Tch.), Graduate Outcomes (GO), Academic Flexibility (AF), Transparency & Accountability (TA), Infrastructure & Resources (IR), Research (Res.), Student Support Services (SSS), International Outlook (IO), Fee & Financial assistance (FFA), Academic Autonomy (AA), and Inclusivity (Inc.). All these parameters along with their overall and group-wise average marking received from the survey are given in Table 1.

Table 1: Overall and Category-wise mean marking scores of eleven HEI's quality parameters in the survey.

$Parameter$	$Overall$	$Category-wise$					
		UG	GS	GR	FA	PR	PA
Tch.	3.2	3.4	3.0	3.3	3.5	3.1	3.2
GO	3.0	3.3	3.0	2.3	2.5	3.3	3.4
AF	3.0	3.0	3.3	3.1	3.0	3.0	2.6
TA	3.0	2.8	2.9	3.2	3.1	2.9	3.0
IR	3.0	3.2	3.0	2.9	2.7	2.8	3.1
Res.	3.0	2.0	2.7	3.4	3.4	3.2	3.0
SSS	2.9	3.2	3.1	2.9	2.4	2.7	3.1
IO	2.8	2.4	2.8	3.1	3.0	2.8	3.0
FFA	2.7	3.1	2.7	3.2	2.0	2.0	3.3
AA	2.5	2.0	2.3	2.7	3.2	2.1	2.6
Inc.	2.5	2.5	2.3	2.8	2.4	2.1	2.7

Survey data was collected in National Capital Region (NCR) of India due to the availability of representative institutions in the NCR. Respondents from Sciences, Medical, Technology, Humanities, and Social Sciences domains of twelve institutions participated in the survey. Seven respondent groups Faculty (FA), Undergraduate (UG), Graduate Study (GS), Graduate Research (GR), Parents (PA), Administrator (AD), and Professional (PR) were identified in the survey. Population in each category except Faculty and Administrator were assumed infinite. Population of faculty in chosen institutions was 5727, whereas no official data was found for the population of administrators [26]. Random sampling with 5% error margin and 95% confidence level were considered for descriptive analysis in the study. Accordingly, 2620 responses were finally considered for analysis in which 438 undergraduates, 463 Graduates, 447 Researchers, 389 professionals, 395 parents, 401 faculty, and 87 administrators, were identified.

The dataset is balanced and satisfies minimum sam-

pling requirements [4, 32, 28, 17, 34, 31]. We use this dataset for clustering purpose in this study. We exclude the administrator category, as its official population was unknown, and the responses obtained from this category were comparatively lesser than that of other categories.

3 Proposed Clustering Method

Unsupervised classification [34, 2, 1, 11] or clustering, partitions a set of data points into different groups such that the data in each group are similar to each other[22, 3, 23]. The proposed clustering method divides the survey data on behalf of frequent marking patterns as these patterns denote behaviors of the respondents [15]. The survey applications generally seek to explore the distinguished preferences of different respondent categories, hence our method divides distinct marking patterns on behalf of their dominance in respondent categories [8, 37]. After dividing patterns in different groups, mean marking of observations corresponding to each subset of the patterns, are calculated. This results in representative marking pattern (mean marking vector) of each respondent category. The mean marking vectors are then matched with original dataset. An observation showing least distance from a mean marking vector is placed in the cluster identified with the index of that mean marking vector. The whole clustering process, employed in this work, can be described in three broad steps: (i) Exploration of distinct marking patterns, (ii) Feature selection, and (iii) Matching. The architecture of the proposed clustering method is depicted in Figure 1.

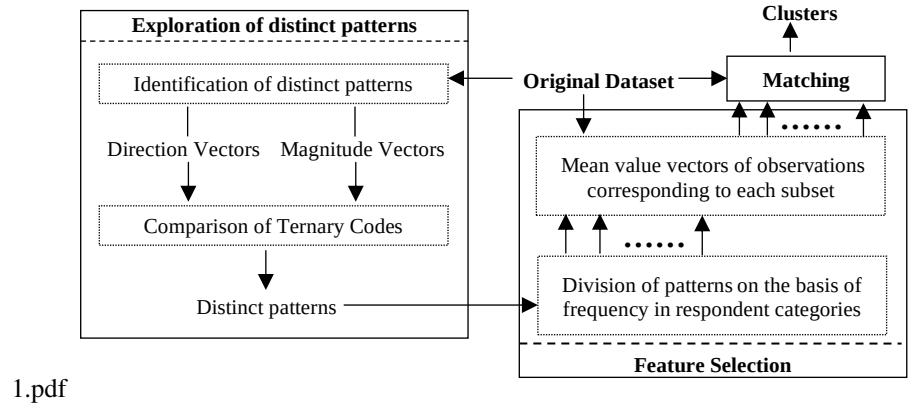
3.1 Exploration of distinct marking patterns

This step identifies distinct marking patterns in the dataset and their corresponding frequencies in each respondent group. This method first calculates two vectors corresponding to each observation: Direction Vector and Magnitude Vector. After that it filters distinct marking patterns on the basis of these vectors. The marking patterns that show unique values for both the vectors are called as *unique* patterns (frequency equals one). Method records these unique patterns and their respondent group. The marking patterns having equal values corresponding to both of these vectors are called as *repeated* patterns (frequency more than one). Method records group labels for each repeated pattern. It counts and records the frequency of each repeated pattern in each of the respondent groups. By this way, the method filters overall distinct patterns and records their frequencies. These patterns along with their frequencies are sent to the next step for further processing.

The proposed method treats survey observation as a single continuous vector started from the left (leftmost is the first variable). The method calculates two vectors: *direction* and *magnitude* regarding each observation. A direction vector records directional information of each variable in the observation. Direction of a variable denotes the relative significance of it with respect to the variable immediately before it. If the value of a variable is greater than the variable preceding it, then this variable denotes upward direction represented by a '$+1$'. It means that this variable is relatively more important to the respondent than the variable preceding it. If the value is less than the value before it, then this shows downward direction which is represented by '-1'. If both the values are same, then it denotes a straight and it is represented by a '0'. In this way three directions are defined for each variable. For the first variable, the direction is defined with respect to the half of the maximum value allowed in survey. Finally, a vector corresponding to each observation is created that contains the direction values for each variable in the observation. This vector is said as the direction vector. Since, a direction vector may contain one of these three equidistant values hence it can be represented by a ternary code and its equivalent decimal value. Figure 2 represents the method of creating direction vector corresponding to an example survey observation that has five variables and '4' as the maximum marking value.

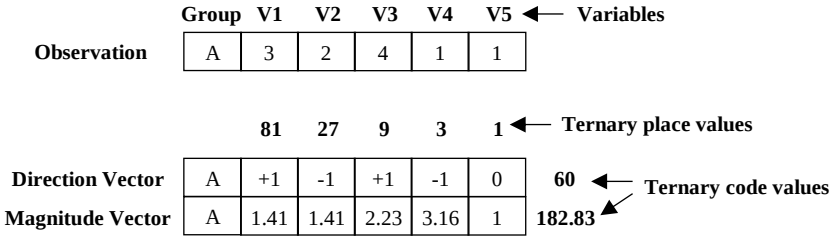
Second vector is called as the magnitude vector. Magnitude vector records magnitude of difference in each variable with respect to the variable preceding it. Magnitude of difference is measured by finding the Pythagorean Distance from its preceding neighbor multiplied by the place value defined by ternary coding. The distance (hypotenuse) is found by Pythagoras formula where a perpendicular equals the difference between variable values, and base equals to one (horizontal distance between two successive variables). For an instance, if two neighboring variables A and B contain the values 2 and 4 respectively, then magnitude of difference for variable B with respect to A will be $\sqrt{(4-2)^2+1^2} = \sqrt{5}$. In this way, the magnitude of difference is recorded for each variable in the magnitude vector. Finally, a ternary code value of the magnitude vector is measured. Figure 2 depicts the whole procedure of calculating direction and magnitude vectors with respect to an example observation.

On behalf of direction and magnitude vectors, the method searches for unique as well as repeating patterns. The observations having unique decimal values for both direction and magnitude vectors, are unique patterns having no repetitions. All such direction vec-



1.pdf

Figure 1: Architecture of the proposed clustering method



2.pdf

Figure 2: Example survey observation with its direction and magnitude vectors

tors along with their group labels are recorded in the list named as distinct patterns. The method now finds repetitive patterns with their frequencies in different groups. Two or more observations having equal decimal values for both direction and magnitude vectors, are repetitive patterns. The direction vector of such observations along with their frequencies in each of the respondent groups are recorded and added to the list of distinct patterns. The purpose of recording the frequencies of a repetitive pattern in different category is to find the category in which the pattern shows its dominance (most frequent). This list of distinct direction vectors – unique as well as repetitive patterns – represents total number of different marking patterns in a survey data.

3.2 Feature Selection

Statistical measures such as mean have the general tendency to suppress delicate differences among patterns having small values. Due to this reason, it is logical to guide the centroids on some basis that could preserve the dissimilarity. The effect of suppression can be reduced if dataset can be divided into subsets on behalf of the features signifying the dissimilarities. According to the stakeholder theory, the grouping of stakeholders is done on the basis of their divergent role, or influence etc. [24, 19]. This signifies that the interests – marking-patterns – of different respondent groups taking part in survey are different. Hence, applying aggregation measures after dividing observations on the basis of respondent categories will produce quite different results than overall aggregation of the dataset. The effects of aggregation over the data can be seen in Table 1. We can see a considerable variation in mean values of parameters inside category-wise columns of Table 1 whereas the overall results shows equal mean values for more than half of the number of parameters. Due to such suppression of dissimilarities by aggregation, a mechanism is needed for survey data that preserves the small differences among observations. Our clustering method thus uses group labels to divide the dataset into smaller subsets and then apply average operation to locate the centroids.

Survey generally seeks to explore the distinct marking patterns of different respondent groups therefore the purpose of the proposed clustering method is to guide the mean centroids in such a way that their position in data-space will preserve distinguished marking patterns

of respondent groups. For achieving this the proposed method divides distinct patterns explored in the previous step on the basis of their occurrence in different respondent groups. A pattern is said to belong a *respondent* group, if its occurrence in that category is most frequent or dominant. Thus, the the proposed method separates out the unique patterns, '*frequency equal to one*', by simply using their group labels. The Repetitive patterns 'emphfrequency more than one', are scrutinized on behalf of frequencies in different respondent groups. Each repetitive pattern is identified with one of the respondent groups in which the pattern has highest frequency of occurrence. In this way, the distinct patterns are divided into the subsets equal to the number of respondent categories. As the proposed method utilizes category labels for identifying the features of different group hence it does not require manual setting of clustering parameters, namely the number of clusters.

Finally, all of the survey observations corresponding to each distinct pattern in every subset of the distinct patterns are filtered out from original dataset. Mean value of each variable in the subset, is then calculated. Resulted vector of mean values is thus the representative marking habit of that respondent group. This procedure is followed for all the subsets of distinct patterns. By this way, the representative mean marking vector of each respondent group is calculated. These mean value vectors are then used as mean centroids for the purpose of clustering. Mean value vector corresponding to a group retain its dominant features as these vectors are calculated according to their dominant frequencies of the occurrence. Finally, the mean value vectors are sent to the next step –matching for making clusters of the dataset.

3.3 Matching

The mean value patterns are now matched with the original dataset. All mean value patterns are first indexed in-order to name the clusters. Indexing is done from one, up to the number of total respondent groups in the dataset. Each observation from the original dataset is then matched with all of the mean value vectors found in the feature selection step. Euclidean Distances of the observation from each mean marking vector are then calculated [10]. An observation under the current consideration is said to belong the cluster corresponding to which it shows smallest Euclidean distance. In this way, all the observations in the dataset are assigned a cluster index.

For further elaboration of this method, let us consider a short sample of survey in which three respondent categories, namely A, B, C are involved. Let the sample contain five observations having three variables. Figure 3 depicts the whole procedure of the proposed clustering over such sample observations. A Likert scale of four levels is assumed for this sample. The method first calculates the direction and magnitude vectors corresponding to each observation. Based on the ternary code values of both direction and magnitude vectors, unique and repetitive patterns are identified. Since, observation 1 and 4 in Figure 3 have unique values regarding both of the vectors hence the direction vectors of both observations 1 and 4 are filtered as unique patterns. Observations 2, 3, and 5 have equal values for direction as well as magnitude vector therefore the direction vectors of these observations are filtered as repetitive pattern. The frequency of the repetitive patterns is recorded, that is, 2 in B category and 1 in C category. As the frequency of repetitive pattern is higher (2) in B as compared to C (1) thus the pattern is assigned the category B. The list of total distinct patterns of the survey is then created which contains a total of three distinct patterns. Now, the observations corresponding to these distinct patterns are fetched. Mean value vector of each category is then calculated as depicted in Figure 3. Finally, the original dataset is matched with these mean centroids (vectors), and three distinct clusters are created, wherein Cluster C1 contains one observation that is observation 1, Cluster C2 contains observation 4, and Cluster C3 contains three observations that are 2, 3, and 5.

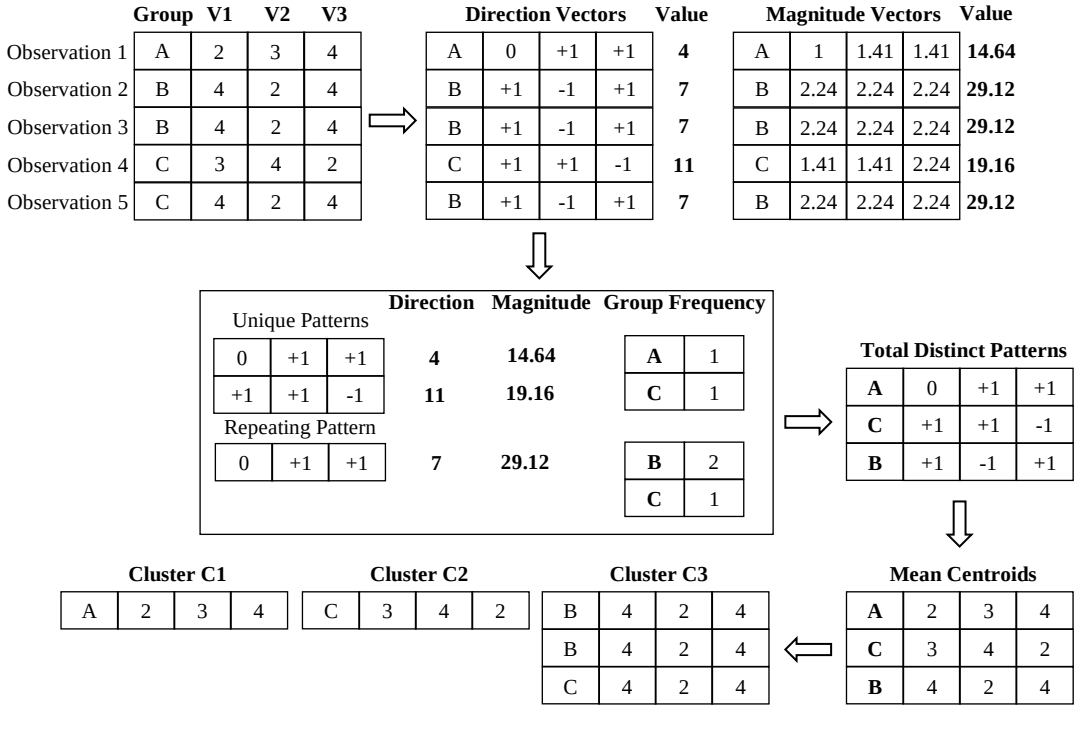
Table 2: Results of the proposed clustering over the survey dataset

$Clusters$	UG	GS	GR	FA	PR	PA
$C1$	19	27	53	275	14	58
$C2$	311	78	7	6	30	31
$C3$	35	111	24	20	15	41
$C4$	16	80	269	20	42	13
$C5$	45	61	73	40	268	44
$C6$	12	106	21	40	26	202
$Total$	438	463	447	401	389	395

4 Analysis of Clustering Results

We apply the proposed clustering over the survey data, the description of which is given in Section 2. In total 2533 responses (excluding administrator) are considered for clustering. The proposed method identifies a total of 755 distinct marking patterns of which frequency ranges from 1 to 35. Method detects six respondent categories in the dataset so it made six clusters named as C1 up to C6. The results of the proposed clustering are given in Table 2.

We can see in Table 2 that each cluster C1 to C6 con-



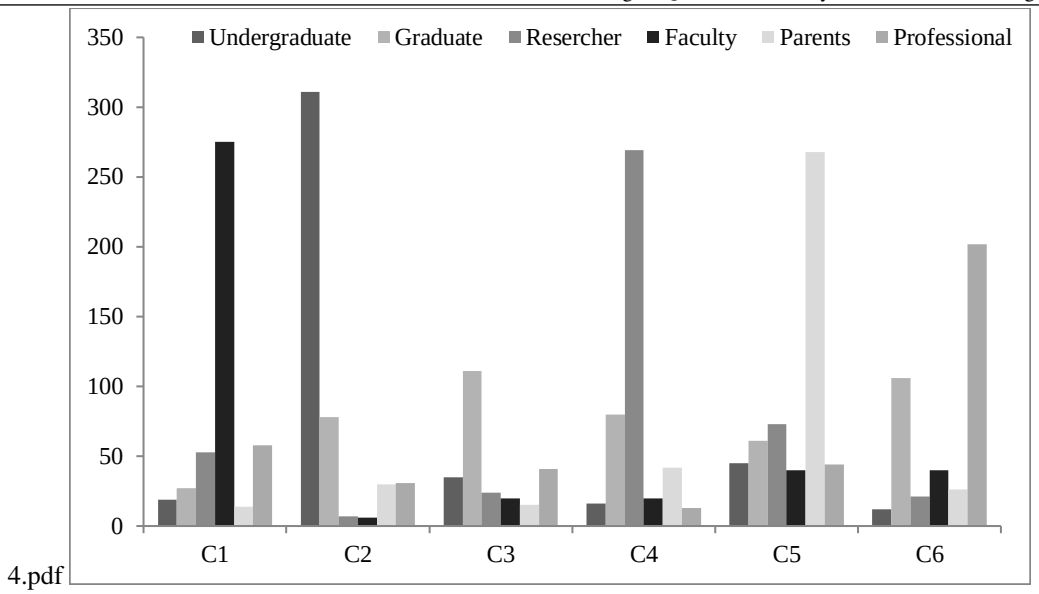
3.pdf

Figure 3: The proposed clustering procedure applied over a sample survey observations

tains a dominant majority of responses from any one of the specific categories. For more elaborated representation of results, we utilize bar-graphs in Figure 4, which represent the population of each category in each of the clusters. According to the population, i.e., height of the bars, we can easily link each cluster with a particular respondent category. For example, Cluster C1 contains a high proportion (62%) of faculty thus this cluster represents marking preferences of faculty. Cluster C2 contains a large population of undergraduates (67%), which suggests that it majorly contains the preferences of undergraduates. Cluster C4 also contains a quite high population (61%) of graduate researchers suggesting that this cluster contains dominant preferences of research scholars. Clusters C5 and C6 contain approximately 50% population of parents and professionals respectively; this also suggests that the preferences of the parents as well as the professionals are specific and distinct from rest of the other communities. A comparatively less dominant majority of graduate study is found inside the cluster C2 (45%) however, this number is also sufficient to signify the dominance of graduates in C2.

In addition to the category-wise population in clusters, we can also observe distribution of each category across the clusters in the columns of Table 2. The majority of the responses in all categories except graduate study (GS) are condensed in one of the specific clusters. This distribution clearly shows that the proposed method is quite capable in segregating the survey responses based on the marking patterns of respondent categories. Somewhat even spread of graduate community in the clusters C3, C4, and C6 suggests that the graduate study is a diverse community in terms of the preferences. Overall with the help of the illustrations of Table 2 and Figure 4, we can conclude that almost all of the academic communities can be differentiated according to their specific marking patterns. These results, hence, satisfy the stakeholder theory, which is well established in higher education domain. Thus, the results prove that clusters made by our method are quite meaningful.

For verifying the strength of the proposed clustering method, we compare it with the K-means clustering method. We apply K-means clustering over the same dataset while taking the value of K as six. The K-means clusters are named as K1 up-to K6. The results of K-means are represented by Table 3 and Figure 5. We can see in the figure that population of faculty (62%) in cluster K1, and population of undergraduates (56%) in cluster K3, are significantly large, thus K1 and



4.pdf

Figure 4: Population of respondent categories inside each cluster made by the proposed method

Table 3: Results of K-means clustering with K=6 for the survey dataset

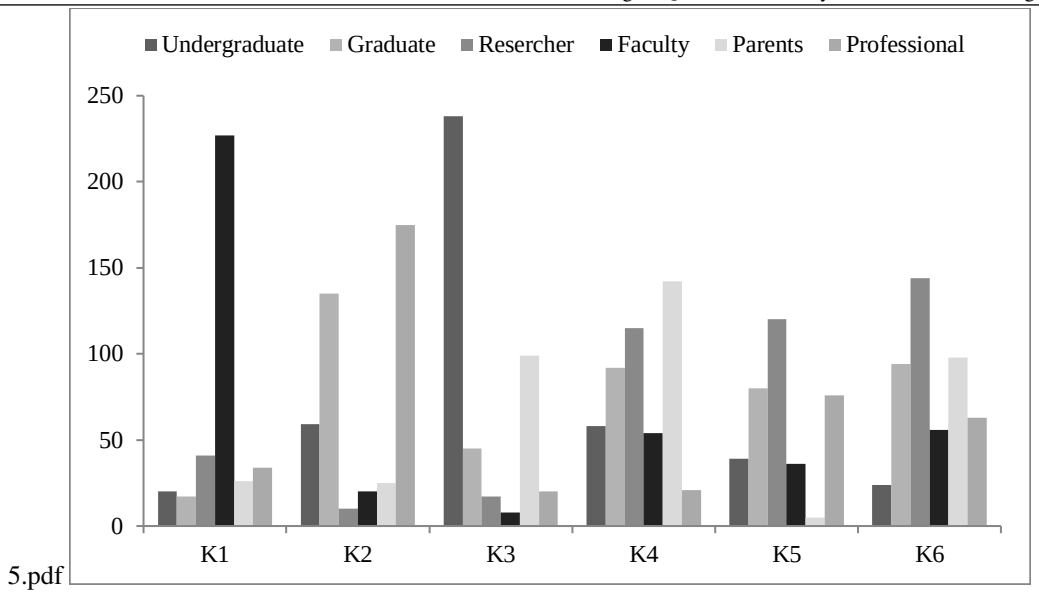
$Clusters$	UG	GS	GR	FA	PR	PA
K1	20	17	41	227	26	34
K2	59	135	10	20	25	175
K3	238	45	17	8	99	20
K4	58	92	115	54	142	21
K5	39	80	120	36	5	76
K6	24	94	144	56	98	63
Total	438	463	447	401	389	395

K2 can be identified with faculty and undergraduates respectively. Besides these two, none of the K-means clusters (K2, K4, K5, and K6) can be linked with any respondent category. This is due to the fact that either the clusters contain majority population from multiple categories, or the categories inside the clusters are evenly distributed. This phenomenon can be seen by observing Figure 5 and Table 3. For example, cluster K2 contains the majority of both graduate study and professionals, whereas the graduate research category is distributed among clusters K4, K5, and K6 with almost equal proportions. Hence, these clusters cannot be linked with any particular category. Thus, overall K-means clustering results are not sufficiently meaningful and do not satisfy the stakeholder theory. This suggests that the proposed clustering method outperforms K-means.

5 Discussion

Cluster analysis of quantitative survey data is an important tool for grouping similar behaviours of the respondents [39]. However, proper clustering of quantitative survey data requires significant consideration due to its distinct properties such as small ordinal values and the associated side information [35]. Most of the clustering methods utilize aggregate statistics that suppress the frequent dissimilarities represented by smaller values. On the other hand marking patterns in survey data denote the behaviors of respondents therefore clustering based on marking pattern is more suitable for survey data. Due to these reasons, a specialized clustering method is required for the survey data that can take care of distinct properties of it and provide meaningful results.

Therefore, we, in this paper, have developed a specialized clustering method for quantitative survey data that combines the features of value based and pattern based clustering methods. The proposed method does not require presetting of clustering parameters as it uses associated information for that purpose. It utilizes pattern based similarity for identifying the distinct behaviors and for guiding the mean centroids at positions that best describes marking habits of the respondents. We have applied the proposed method over a real survey dataset. Clustering results of the proposed method are quite meaningful and interpretable. Moreover, the proposed method outperforms K-means with respect to its suitability for quantitative survey data. The study done



5.pdf

Figure 5: Population of respondent categories inside each cluster made by K-means (K = 6)

in this paper is important in context of future research in the field of survey data analysis. For example, more effective parameters and distance measures can be formulated for survey data in future that work purely on behalf of the pattern based similarity.

6 Conclusion

We have proposed a specialized clustering method for quantitative survey data that combines best features of both value based and pattern based similarity. The proposed method converts survey observations into patterns, and then filters these patterns on the basis of their frequencies in different respondent groups. As it utilizes group labels for estimating clustering parameters and selecting representative features from data therefore it does not require pre-setting of clustering parameters. We applied the proposed method over a real survey dataset and compare its results with K-means clustering with respect to the benchmark stakeholder theory. The results of the proposed clustering were quite meaningful and satisfied the stakeholder theory whereas the results of K-means method were not fully interpretable, and did not satisfy stakeholder theory. This phenomenon suggests that the proposed method is quite suitable for quantitative survey data. This paper presents a foundation for designing dedicated analysis techniques for quantitative survey data. Future research will be focused on designing robust pattern based clustering for survey data.

References

[1] Affonso, E. T., Nunes, R. D., Rosa, R. L., Pivaro, G. F., and Rodríguez, D. Z. Speech quality assessment in wireless voip communication using deep belief network. *IEEE Access*, 6:77022–77032, 2018.

[2] Affonso, E. T., Rosa, R. L., and Rodríguez, D. Z. Speech quality assessment over lossy transmission channels using deep belief networks. *IEEE Signal Processing Letters*, 25(1):70–74, 2018.

[3] Amine, A., Elberrichi, Z., Simonet, M., and Malki, M. Evaluation and comparison of concept based and n-grams based text clustering using som. *INFOCOMP Journal of Computer Science*, 7(1):27–35, 2008.

[4] Bluman, A. G. *Elementary Statistics: A Step by Step Approach*. McGraw-Hill Higher Education New York, NY, 2009.

[5] Burrows, A. and Harvey, L. Defining quality in higher education: the stakeholder approach. In *Proc. AETT Conf. Quality in Education*, pages 6–8, 1992.

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

[7] Cheng, Y. and Church, G. M. Biclustering of expression data. In *Proc. ISMB*, volume 8, pages 93–103, 2000.

[8] Church, A., Waclawski, J., and Kraut, A. *Designing and Using Organizational Surveys: A Seven-Step Process*. Business and Management Series. Wiley, 2001.

[9] CUG. University League Tables-2018. https://www.thecompleteuniversityguide.co.uk.

[10] D'Agostino, M. and Dardanoni, V. What's so special about Euclidean distance? *Social Choice and Welfare*, 33(2):211–233, 2009.

[11] de Almeida, F. L., Rosa, R. L., and Rodríguez, D. Z. Voice quality assessment in communication services using deep learning. In *2018 15th International Symposium on Wireless Communication Systems (ISWCS)*, pages 1–6, 2018.

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

[13] Freeman, R. *Strategic Management: A Stakeholder Approach*. Pitman series in business and public policy. Cambridge University Press, 2010.

[14] Grice, J. W. Observation oriented modeling: preparing students for research in the 21st century. *Comprehensive Psychology*, 3:05–08, 2014.

[15] Grice, J. W. From means and variances to persons and patterns. *Frontiers in Psychology*, 6:1007, 2015.

[16] GuimarÃ£es, R., Rodríguez, D. Z., Rosa, R. L., and Bressan, G. Recommendation system using sentiment analysis considering the polarity of the adverb. In *2016 IEEE International Symposium on Consumer Electronics (ISCE)*, pages 71–72, 2016.

[17] GuimarÃ£es, R. G., Rosa, R. L., De Gaetano, D., Rodríguez, D. Z., and Bressan, G. Age groups classification in social network using deep learning. *IEEE Access*, 5:10805–10816, 2017.

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

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

[20] Kriegel, H.-P., Kröger, P., Sander, J., and Zimek, A. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining & Knowledge Discovery*, 1(3):231–240, 2011.

[21] Kriegel, H.-P., Kröger, P., and Zimek, A. 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, 2009.

[22] Kumar, V., Chhabra, J. K., and Kumar, D. Performance evaluation of distance metrics in the clustering algorithms. *INFOCOMP Journal of Computer Science*, 13(1):38–52, 2014.

[23] Lasmar, E. L., de Paula, F. O., Rosa, R. L., AbrahÃ£o, J. I., and Rodríguez, D. Z. Rsrs: Ridesharing recommendation system based on social networks to improve the userâs qoe. *IEEE Transactions on Intelligent Transportation Systems*, 20(12):4728–4740, 2019.

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

[25] Mitchell, R. K., Agle, B. R., and Wood, D. J. 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, 1997.

[26] NIRF. National Institutional Ranking Framework (NIRF), Ministry of Education, Government of India. https://www.nirfindia.org/Home, 2018.

[27] QS. World University Ranking-2018. https://www.topuniversities.com, 2018.

[28] Rodríguez, D. Z., Abrahao, J., Begazo, D. C., Rosa, R. L., and Bressan, G. Quality metric to assess video streaming service over tcp considering temporal location of pauses. *IEEE Transactions on Consumer Electronics*, 58(3):985–992, 2012.

[29] Rodríguez, D. Z., Rosa, R. L., and Alfaia, E. C. A simple method to measure the image complexity on a fault tolerant cluster computing. In *2010 Sixth*

Advanced International Conference on Telecommunications, pages 549–554, 2010.

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

[31] Rosa, R. L., Rodríguez, D. Z., and Bressan, G. Sentimeter-br: A social web analysis tool to discover consumers' sentiment. In *2013 IEEE 14th International Conference on Mobile Data Management*, volume 2, pages 122–124, 2013.

[32] Rosa, R. L., Rodríguez, D. Z., and Bressan, G. Music recommendation system based on user's sentiments extracted from social networks. *IEEE Transactions on Consumer Electronics*, 61(3):359–367, 2015.

[33] Rosa, R. L., Rodríguez, D. Z., Schwartz, G. M., de Campos Ribeiro, I., and Bressan, G. Monitoring system for potential users with depression using sentiment analysis. In *2016 IEEE International Conference on Consumer Electronics (ICCE)*, pages 381–382, 2016.

[34] Rosa, R. L., Schwartz, G. M., Ruggiero, W. V., and Rodríguez, D. Z. A knowledge-based recommendation system that includes sentiment analysis and deep learning. *IEEE Transactions on Industrial Informatics*, 15(4):2124–2135, 2019.

[35] Sadh, R. and Kumar, R. Clustering of quantitative survey data: A subsystem of EDM framework. In Singh, V., Asari, V. K., Kumar, S., and Patel, R. B., editors, *Computational Methods and Data Engineering*, pages 307–319. Springer, 2020.

[36] SRC. Academic Ranking of World Universities. http://www.shanghairanking.com, 2018.

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

[38] THE. World University Rankings-2018. https://www.timeshighereducation.com.

[39] van der Hoef, H. and Warrens, M. J. Understanding information theoretic measures for comparing clusterings. *Behaviormetrika*, 46(2):353–370, 2019.

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

[41] Wang, H., Wang, W., Yang, J., and Yu, P. S. Clustering by pattern similarity in large data sets. In *Proc. ACM SIGMOD Int. Conf. Management of Data*, pages 394–405, 2002.

[42] Xu, D. and Tian, Y. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193, 2015.

[43] Zegarra Rodríguez, D., Rosa, R. L., and Bressan, G. A proposed video complexity measurement method to be used in a cluster computing. In *2013 IEEE Global High Tech Congress on Electronics*, pages 76–77, 2013.