Unification of Numerical and Ordinal Survey Data for Clustering-based Inferencing

BHUPENDERA KUMAR[1]
RAJEEV KUMAR[2]

Data to Knowledge (D2K) Lab
School of Computer & Systems Sciences
Jawaharlal Nehru University, New Delhi 110 067, India
[1]bkchauhan86@gmail.com, ORCID: 0000-0001-7703-0417
[2]rajeevkumar.cse@gmail.com, ORCID: 0000-0003-0233-6563

Abstract. With the proliferation of surveys for almost every issue governing our life with various parameters and a variety of data, it becomes necessary for a researcher to unify these data followed for extracting inferences from the survey. Data from quantitative surveys are clustered to reveal respondents' divergent and dominant tendencies. It aims to investigate the general trends among the respondents' categories. Due to the unique characteristics of survey data, popular clustering techniques based on value similarity are inadequate. In this paper, we attempt to unify the numerical data with the ordinal data of a survey. We model the data with a Gaussian distribution; therefore, we first convert the numerical data to ordinal data following the distribution; this may be the governing attribute for deciding the clusters. Then, we use K-means clustering with varying numbers of clusters. We implement the proposed methodologies on real survey data and compare the clustering efficiency. More crucially, it appropriately uses the ordinal attributes order information and numerical attribute statistical information for clustering. Extensive testing demonstrates that the suggested unification works better on real data sets than its contemporaries.

Keywords: Gaussian Distribution, Clustering, Numerical Data, Ordinal Data, Unification.

(Received May 12th, 2023 / Accepted June 1st, 2023)

1 Introduction

Entrepreneurs have shaped the world in every aspect. We thought about attempting to comprehend the fundamental characteristics of what makes a great entrepreneur. Numerous opportunities are presented by entrepreneurship to the entrepreneur as well as to society. Entrepreneurs are transforming the world; a universe venture needs an environment to sustain it. Since a shortage of entrepreneurs can have a direct detrimental influence on the economy and slow down the development agenda in the nation, the lack of entrepreneurial skills and desire among university students raises worrying concerns. For startup founders and the general public, entrepreneurship offers up a world of opportunities and doors, and we aim to use machine learning algorithms to address the issue of the lack of entrepreneurial ability [30, 33]. With certain ages of university students, their interests can be grouped into educational backgrounds towards getting income through entrepreneurship [27, 6, 32]. This process may be done through clustering [19].

A given clustering method's usefulness depends on the viewers' perspectives [24, 31]. The appropriateness of the qualities of the data will determine the clustering strategy. For non-convex data, partition-based clustering techniques like K-means are inappropriate [1]. DBSCAN and other distribution-based clustering techniques are inappropriate for sparse collections with dif-

ferent densities. When comparing values, most of these clustering techniques use aggregate statistics, such as mean, variance, and others, deemed improper for behavioral research. The poll results have unique characteristics, like a fixed limited range of ordinal values and related data presented in the form labels for groups [10, 21, 3].

Most value-based clustering techniques are inappropriate for survey data because of these characteristics. On the other hand, pattern-based clustering is relevant in these scenarios as patterns in survey data represent the marking habits or behavior of the respondents [26]. Value-based clustering is preferable due to its simplicity and processing efficiency. The current pattern-based clustering techniques are created for particular applications and data types, such as microarray analysis of gene expression data [11]. Therefore, the quantitative survey data cannot be clustered using the current pattern clustering algorithms [20]. These factors need a particular clustering technique for correctly segregating quantitative survey data [4]. However, the simultaneous gathering and analysis of numerical and ordinal data in a single study make it a tedious task to find skilled entrepreneurs through different clustering techniques. Thus, in this paper, we apply the Gaussian distribution policy to unify these data. We compare the results of this unification through K-means clustering.

The rest of the paper is organized as follows. Section 2 gives a brief description of related work. The proposed method is explained in Section 3. The used dataset is described in Section 4.1. Analysis of the clustering results is included in Section 4.2. The discussion of the proposed method, along with the future work, is described in Section 4.3. Section 5, finally, concludes the findings of this work.

2 Related Work

Here, we take a more systematic approach and concentrate on issues such as entrepreneurs' category, pattern-based, and arbitrarily oriented clustering. Our methodical perspective is based not on the application situations but rather on the inherent methodological variations among the numerous families of techniques based on diverse spatial intuitions. As a result, we work to incorporate methodologies' fundamentally diverse points of view into our intuition for patterns in the data.

2.1 Stakeholder and Entrepreneur in Higher Education

Universities worldwide must carefully reevaluate their place in society and their interactions with many constituencies, stakeholders, entrepreneurs, and communities. Using entrepreneur analysis as a tool to help colleges categorize entrepreneurs and ascertain the significance of each group. Universities are increasingly expected to take on the third mission and interact with local and industrial partners. While government initiatives and incentive programs aim to motivate universities to engage with the outside world more, there are still some significant obstacles to these connections. Universities must carefully choose their stakeholders and determine the correct degree of differentiation to fulfill their responsibility towards entrepreneurship [12]. The internal and external stakeholders' impressions of scenarios for the quality assurance of stakeholder interactions are based on interviews and expert panel data. It focuses mainly on examining how institutions might balance the perspectives of internal and external stakeholders for quality assurance. The findings demonstrate that developing adaptable quality assurance procedures that balance the institutions' academic and entrepreneurial goals is a crucial problem for higher education institutions [18].A method of monitoring people experiencing psychological disturbances, particularly stress. The sentences taken from social networks are mood-filtered and graded using a sentiment-analysis methodology that considers factors like gender and age. The solution alerts authorized parties of the users' emotional changes. This may help to depict the behavior of an entrepreneur [28].

2.2 Pattern and Direction Based Clustering in High Dimensions

The goal of clustering is to divide datasets into groups, where objects within the same group are similar to one another in terms of a particular similarity, and objects within separate groups are dissimilar. Even though clustering is generally a rather dignified problem, new strategies have been proposed to address the challenges posed by using modern automatic data generation and acquisition capabilities in increasing applications, which produce a large amount of high-dimensional data [14, 16]. The use of distance measurements is crucial in cluster analysis. Only so many distance metrics work well for all kinds of clustering issues. Kumar et al. measured ten distinct distances tested against eight different clustering techniques. Three criteria were used to calculate the distance measurements' quality: accuracy, inter-cluster, and intra-cluster distances. The nature of the data and the clustering methods affect the effectiveness and quality of various distance measurements [17]. The spectral technique performed exceedingly well when the

adopted methods were used in their default configurations. The *pCluster* paradigm finds use in various contexts, including e-commerce applications like collaborative filtering and the handling of scientific data like the DNA microarray [36]. A brand-new measurement called the Video Complexity Index (VCI) takes both the spatial and temporal video properties into account [25].

Data from quantitative surveys are clustered to reveal respondents' divergent and dominant tendencies. It aims to investigate the general trends among the respondents' categories. Popular clustering techniques based on the similarity of values are inadequate for survey data because of their unique characteristics. Separating the replies based on marking patterns is an efficient way to determine the prevailing behaviors since marking patterns in survey data represent respondent behavior. To produce meaningful results, quantitative survey data must combine the advantages of both value-based and pattern-based approaches [29, 15].

2.3 Gaussian Distribution with K-means Shift

Using mixtures of multivariate normal inverse Gaussian (MNIG) distributions, skewness, and heavy tails in data can be detected and grouped. The number of components must either be known a priori or calculated a posteriori using a model selection criterion after generating results for a range of potential component numbers to perform cluster analysis using a typical finite mixture model framework. Different model selection criteria, however, can cause different numbers of components to produce uncertainty. For the mixes of MNIG distributions, the Dirichlet process mixture model is suggested [5]. A general technique for viewing the outcomes of model-based clustering in a Gaussian approach was presented by Biernacki et al. Because it solely relies on the distribution of classification probabilities, this method enables the display of any model-based clustering performed on any data. It allows for the interpretation of model-based clustering findings but not for choosing the most effective clustering technique (choosing a clustering method has to be performed before through a classical model selection process). This makes it different from the exploratory visualization methods that are frequently used [2]. The best-known clustering algorithms, such as iterative (K-means, random swap, expectation-maximization), hierarchical (pairwise nearest neighbor, split, split-and-merge), evolutionary (genetic algorithm), neural (self-organizing map), and fuzzy (fuzzy C-means) approaches, are used in the parametric Gaussian mixture model (GMM) and non-parametric vector quantization (VQ) models [13].

A non-parametric, iterative method called the mean shift (MS) algorithm has been used to identify modes in an estimated probability density function (pdf). Youness Aliyari and Ghassabeh established the sequence convergence produced by the MS algorithm for an estimated pdf with isolated stationary spots. They provide an applicable requirement for secluded inactive areas in the Gaussian kernel-estimated pdf [7]. An established method for categorizing multivariate observations is K-means clustering. To understand which variables define clusters, it is observed that the resulting centroid matrix of clusters by variables. However, the centroid matrix may only sometimes accurately represent between-cluster differences. By reducing the total within-cluster deviation of the n partitions in the hierarchy, HMC generates a set of nested partitions using a centroid-based model estimated by least squares [35].

2.4 Unification of Data

Emphasis on the difficulties of unification rather than its advantages, with particular attention paid to long-term economic growth and the practicalities of societal and political integration. The extent to which word choice cognition of unification is still unknown [23]. There are many categorical data types, including text data, DNA sequences, and Census Bureau data. While such data are simple for humans, many classification systems, like support vector machines and others, cannot directly use them because they require the underlying data to be expressed numerically. Most learning techniques today transform categorical data into binary values, which could lead to excessive dimensionality and sparsity. CNFL employs eigendecomposition to transform the closeness matrix into a lesser space, which may be used to represent examples for classification or clustering. It first uses simple matching to determine proximity between instances [9]. It is easy to pose useless queries during any data analysis. Understanding data scaling might occasionally make it easier to spot gibberish, but we must use the proper logic.

We begin with the data and our hypotheses about the circumstances underlying the data rather than basing the choice of statistical procedures on the scale type. What we seek to discover from the data informs how we conduct the data analysis [34]. Two common cluster analysis issues exist in selecting the group sizes and the scale invariance; Giordan and Diana proposed a novel clustering technique explicitly designed for ordinal data. A multinomial model, a cluster tree, and a pruning approach are used to group the objects. Using simulations, two types of pruning are examined [8]. The decision tree algorithm supports categorical and

numerical variables by calculating a target based on the defined rules. This property is used to propose a new hybrid model that combines a decision tree with a different regression technique to assess mixed data. GA, which optimizes the new cost function, is the algorithm utilized to produce accurate clustering results. We may compare and determine if a GA-based clustering algorithm is workable for high dimensional data collections with mixed features [22]. An innovative distance metric that preserves the order link between ordinal values while measuring the intra-attribute distances of nominal and ordinal characteristics in a unified manner. An entropy-based distance metric for ordinal attributes was devised to estimate the distance between categories of an ordinal attribute, which uses the underlying order information. The next step is to generalize this distance measure and suggest a single one that applies to ordinal- and nominal-attribute categorical data [37].

3 The Proposed Method

A probability distribution that is symmetric about the mean is the normal distribution. It is also called the Gaussian distribution. It demonstrates that data close to the mean occur more frequently than data far from the mean. Its probability density function's general form is:-

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right),$$

where the parameter μ is the mean, while the parameter σ is its standard deviation.

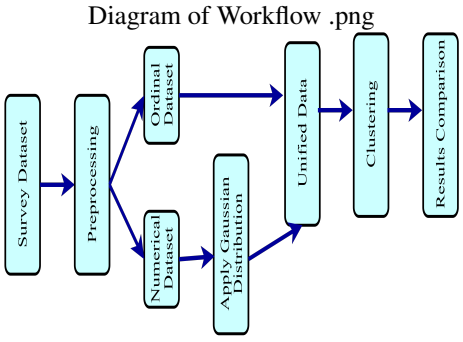


Figure 1: Block Diagram of the Workflow.

When the survey data are collected, more than one data type may be collected. These may be ordinal, numerical, or nominal. All types of data have an impact on survey research. If we can include all data types and deduce our results, they will be more authentic and beneficial. Here in this work, we consider two data types among many more which are numerical and ordinal. Now come to further inferences from these datasets, we unify them. We will map numerical data on ordinal data to get better results. So in this process, we distribute the numerical data according to Gaussian distribution. The whole process of the work is shown in Figure 1 and described as follow:-

1. **Preprocessing:-** The collected survey dataset through a questionnaire is filtered. The data is segregated into two parts.
   (a) Numerical valued attributes
   (b) Ordinal valued attributes
2. **Gaussian Distribution:-** The numerical data is converted into ordinal data, which follows the Gaussian distribution. The numerical data may be (Age in this work) or may not be (Income in this work) Gaussian distributed in the original survey. If the distribution is not Gaussian, arrange the particular attribute's minimum to maximum range in non-decreasing order. Decide the separable criteria for each range (in the original survey) and count the number of instances accordingly. A recursive procedure called to find the more appropriate combination of distribution towards Gaussian distribution (for ordinal data conversion).
3. **Unification:-** Gaussian distributed ordinal data obtained from numerical data is merged with ordinal data.
4. **Clustering:-** K-means is applied to the original and new ordinal datasets. In this process, four to nine clusters are made.
5. **Comparison:-** The comparison is made on efficiency (correctness) and the number of instances in clusters (compactness).

4 Results and Analysis

4.1 The Dataset

In this work, we have used the dataset collected from a survey. This dataset is compiled from a university students' survey on acquiring a habit of an entrepreneur with their Income. This dataset is collected for research purposes. There are 219 responses in this dataset. The distribution of respondents under EducationSector is given in Table 1:-

This dataset comprises ten features in the form of attributes, i.e., Age (A1), Perseverance (A2), DesireToTakeInitiative (A3), Competitiveness (A4), SelfRe-

Table 1: No of Respondents in each category

EducationSector	No of Respondents
Art, Music or Design	21
Economic Sciences, Business Studies, Commerce and Law	32
Engineering Sciences	123
Humanities and Social Sciences	5
Language and Cultural Studies	1
Mathematics or Natural Sciences	4
Medicine, Health Sciences	10
Others	20
Teaching Degree (e.g., B.Ed)	3
Total	219

liance (A4), StrongNeedToAchieve (A6), SelfConfidence (A7), GoodPhysicalHealth (A8), Income (A9), and EducationSector. Age and Income are numerical data. Perseverance, DesireToTakeInitiative, Competitiveness, SelfReliance, StrongNeedToAchieve, SelfConfidence, and GoodPhysicalHealth are in ordinal data. EducationSector is categorical data. The abbreviations used in this work are given in Table 2. All these attributes, along with their overall and attributewise mean received from the survey, is given in Table 3.

Before going to the result section, we ensure the appropriate modeling of the data and the impacts of this process on the clustering results. The following steps are taken:

1. **Data Preprocessing / Feature Extraction:** It includes handling missing values and selecting numerical and ordinal values only for the whole procedure. (First step of proposed method)

2. **Normalization and Standardization:** It brings the data to a common scale. (Second and third step of proposed method)

3. **Model Selection:** Because of the nature of the data (Numerical and ordinal), K-means is best suitable for clustering analysis. (Fourth step in proposed method)

4. **Evaluation and Validation:** After applying the K- means clustering algorithm, the results must be evaluated and validated. Internal validation measures such as silhouette score or Dunn index can be used to assess the clustering quality. But we have used external validation. We use the stakeholder theory [29]; for ground truth, we use Table 1.

Table 2: Abbreviation used for EducationSector

EducationSector	Abbreviation
Art, Music or Design	AMD
Economic Sciences, Business Studies, Commerce and Law	ESBSCL
Engineering Sciences	EC
Humanities and Social Sciences	HSS
Language and Cultural Studies	LCS
Mathematics or Natural Sciences	MNS
Medicine, Health Sciences	MHS
Others	OT
Teaching Degree (e.g., B.Ed)	TD

4.2 Results

In this survey, we have two numerical valued attributes. One is Age, and another one is income. We apply the proposed Gaussian distribution to these two attributes. It is important to note that our goal is to show that clustering may be used more broadly and to evaluate the effectiveness of this strategy when we include a Gaussian distribution compared to a statistical strategy based on distance metrics. We divide our analysis into two parts.

In the first part, we unified the Age and Income attributes. Figure 2 shows the age unification, and Figure 3 shows the income unification. The original age attribute is not fully normally distributed, and the original income attribute is very random in nature because of its uniqueness. After applying the Gaussian distribution method, we can easily map these data on ordinal data.

In the second part, we find the impact of mixing the proposed method in terms of efficiency. Here, we took the attributes one by one. And mix them with the remaining dataset. Figure 4 shows the comparison of the efficiency of clustering. We use the K-means algorithm for clustering. First, we took only age attribute and compared it with the original dataset after clustering. Then, we add income attributes and compare the results with the original dataset again.

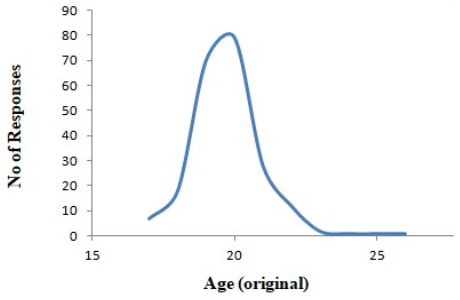
For the analysis of the proposed methodology, we have used performance metrics which is the stakeholder's theory [29], as we have already mentioned earlier in the dataset subsection. This theory validates our assumption that the number of respondents corresponding to their EducationSector should be in the same cluster. We formulate the efficiency of the clustering approach as follows:

$$\eta = \frac{N_C}{N_T},$$

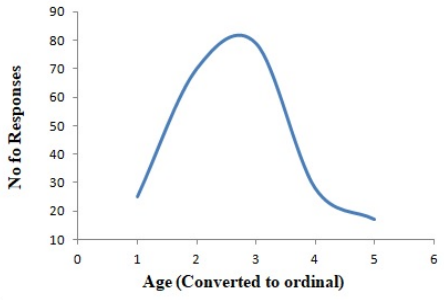
where η is the efficiency, N_C are the number of correctly clustered instances and N_T are the total number

Table 3: Overall and attribute-wise mean

EducationSector	Attributes								
	A1	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**
AMD	20.33	3.19	3.38	3.43	3.57	3.76	3.67	3.38	58799.57
ESBSCL	19.56	3.38	3.72	3.47	3.75	4.09	3.56	3.63	73727.63
ES	19.74	3.38	3.72	3.72	3.81	4.02	3.62	3.61	71495.71
HSS	19.60	3.40	3.60	3.00	4.00	3.80	3.60	3.60	81794.4
LCS	19.00	3.00	5.00	3.00	3.00	5.00	5.00	2.00	50601
MNS	18.75	3.00	3.25	3.25	2.25	3.00	3.25	3.75	65225.5
MHS	19.60	3.40	3.20	3.40	3.90	3.60	3.50	3.70	91045.4
OT	20.00	3.25	3.35	3.45	3.45	3.35	3.30	3.30	72838.8
TD	19.00	4.00	3.67	3.67	3.67	4.00	3.33	3.67	55957.00
Overall	19.75	3.35	3.62	3.59	3.72	3.91	3.58	3.56	71432.07



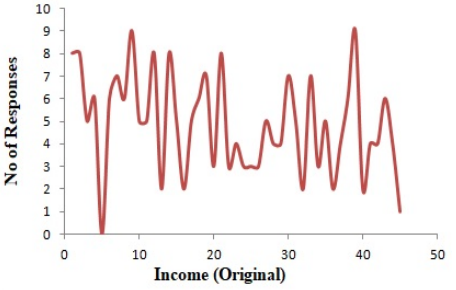
(a) Original Distributed



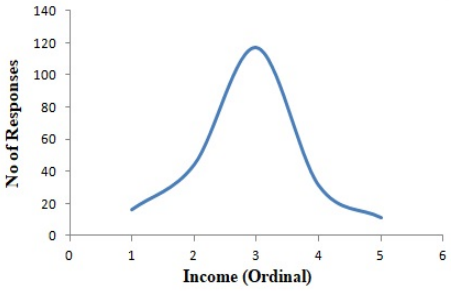
converted$_{d}ata5.jpg$

(b) Normal Distribution

Figure 2: Conversion Comparison of Age attribute on 5-point Likert scale



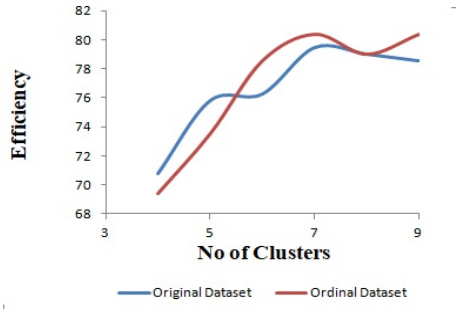
(a) Original Distributed



(b) Normal Distribution

Figure 3: Conversion Comparison of Income attribute on 5-point Likert scale

of instances.



(a) Efficiency Without Income Attribute

(b) Efficiency With Income Attribute

Figure 4: Efficiency Comparison on Different Clusters

4.3 Discussion

An essential tool for classifying respondents' behaviors according to similarity is the study of quantitative survey data. The distinctive qualities of quantitative survey data, such as its small ordinal values, mixed values, and related site information, necessitate careful study for accurate clustering. Most clustering techniques use aggregate statistics, which mask frequent differences reflected by smaller values. Since the marking patterns in survey data represent the responses' actions, clustering according to the marking pattern is more appropriate for survey data. Although numerous recent attempts have been made to suggest clustering techniques for mixed data, few studies have focused explicitly on numerical variables. With this study, we want to overcome this restriction by presenting a particular Gaussian distribution method to deal with the numerical data types prevalent in many contexts, such as questionnaires where the responses are almost always given as numerical values like age or income. The suggested approach is based on a probabilistic model that can be advantageous for inferential reasoning. Through simulations, this point has been researched.

In this work, we are focused on numerical data collected during the survey through a questionnaire. Figure 4 shows that as the number of clusters increased, the correctness and the compactness improved. The efficiency of a cluster of putting instances into a cluster has been improved. First, we did the clustering according to Gaussian distribution without income attributes and added only age attributes. Secondly, we add income attributes too, to find the impact. The results of improved efficiency with income attributes have been significantly enhanced.

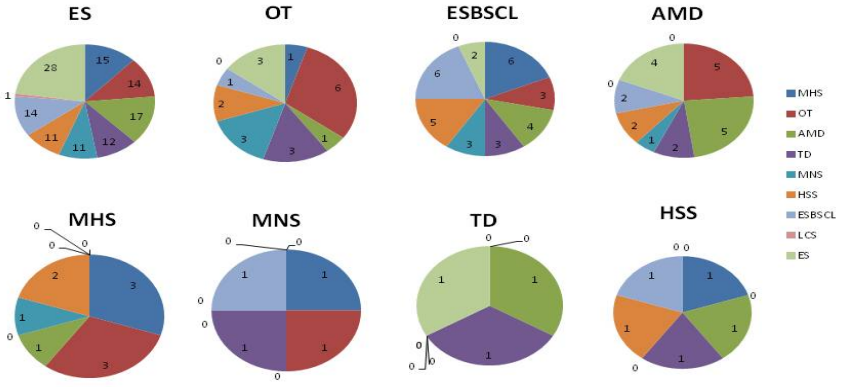
Figure 5 shows the number of respondents present in each cluster. Here, we took a particular case of nine clusters. In LCS of EducationSector, we have only one respondent's response. We have not taken that LCS cluster and shown eight other clusters. In the original dataset, we did not convert numerical data to ordinal data, and the participation of each type of respondent in packing is shown. The ordinal dataset uses the full dataset with ordinal values for each attribute.

4.4 Limitations

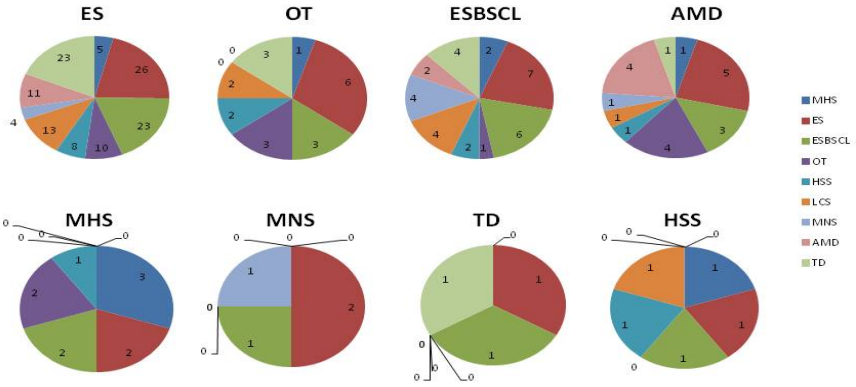
When discussing numerical data, we imply information that can be measured and expressed as numbers. The analyzed data shows these values could be either whole numbers or real numbers. The ordinal data, which includes rating scales and survey results, shows a hierarchy or order of categories or variables. When transforming numerical data into ordinal data, we use various methods to spot patterns, connections, and trends in the data set. These strategies make meaningful analyses and comparisons between various data points possible. When working with infinitely large or real numbers of large precision, it may not be easy to quantify or translate into meaningful ordinal values for analysis.

5 Conclusion

We have presented a customized Gaussian distribution approach for quantitative survey data, mainly for numerical data, to integrate and combine with ordinal data. The dataset is then clustered using the best value-based similarity features. The suggested method transforms ordinal observations from numerical survey data into clusters of respondents. For various clusters, we used K-means clustering. We utilized the proposed technique using actual survey data. We examine the outcomes obtained using the suggested strategy before and after applying it to K-means clustering. The results of the proposed method were reasonably significant and supported the compactness and correctness. According to this occurrence, the suggested strategy is suitable for



(a) With original Dataset



(b) With ordinal Dataset

Figure 5: A Sample of Clustering with Nine Clusters

using quantitative survey data. The foundation for developing specific analytic methods for quantitative survey data is presented in this study. We found that the proposed method adequately gives more accurate and similar instances within each cluster, thus simplifying the inferential process.

In the future, we will include nominal and continuous data to unify survey data. Nominal and continuous data can capture more accurate and detailed information, allowing for more nuanced analysis and modeling. Nominal data also offers essential insights into demographic traits. Thus, nominal and continuous data improve the thoroughness and quality of survey data analysis. By combining numerical, ordinal, nominal, and continuous data, this is feasible to do data analyses that are more reliable and obtainable.

Acknowledgement

The authors thank the editor and reviewer(s) for their valuable comments.

References

[1] 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.

[2] Biernacki, C., Marbac, M., and Vandewalle, V. Gaussian-based visualization of gaussian and non-gaussian-based clustering. *Journal of Classification*, 38(1):129–157, 2021.

[3] Carrillo, D., Nguyen, L. D., Nardelli, P. H., Pournaras, E., Morita, P., Rodríguez, D. Z., Dzaferagic, M., Siljak, H., Jung, A., Hébert-Dufresne, L., et al. Corrigendum: Containing future epidemics with trustworthy federated systems for ubiquitous warning and response. *Frontiers in Communications and Networks*, 2:721971, 2021.

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

[5] Fang, Y., Karlis, D., and Subedi, S. Infinite mixtures of multivariate normal-inverse gaussian distributions for clustering of skewed data. *Journal of Classification*, pages 1–43, 2022.

[6] Ferreira, J. P. B., Junior, F. L., Rosa, R. L., and Rodríguez, D. Z. Evaluation of sentiment and affectivity analysis in a blog recommendation system. In *Proceedings of the XVI Brazilian Symposium on Human Factors in Computing Systems*, pages 1–9, 2017.

[7] Ghassabeh, Y. A. A sufficient condition for the convergence of the mean shift algorithm with gaussian kernel. *Journal of Multivariate Analysis*, 135:1–10, 2015.

[8] Giordan, M. and Diana, G. A clustering method for categorical ordinal data. *Communications in StatisticsâTheory & Methods*, 40(7):1315–1334, 2011.

[9] Golinko, E., Sonderman, T., and Zhu, X. CNFL: categorical to numerical feature learning for clustering and classification. In *Proc. IEEE 2nd Int. Conf. Data Science in Cyberspace*, pages 585–594. IEEE, 2017.

[10] Harvey, L. The new collegialism: improvement with accountability. *Tertiary Education & Management*, 1(2):153–160, 1995.

[11] 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.

[12] 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.

[13] Kinnunen, T., Sidoroff, I., Tuononen, M., and Fränti, P. Comparison of clustering methods: A case study of text-independent speaker modeling. *Pattern Recognition Letters*, 32(13):1604–1617, 2011.

[14] 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.

[15] Kumar, B. and Kumar, R. Difference-attribute-based clustering for ordinal survey data. In *Proc. Int. Conf. Signal Processing & Integrated Networks*, pages 17–27. Springer, 2022.

[16] Kumar, B. and Kumar, R. Entropy-based clustering for subspace pattern discovery in ordinal survey data. In *Proc. Int. Conf. Frontiers of Intelligent Computing: Theory and Applications*, pages 509–519. Springer, 2022.

[17] 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.

[18] 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.

[19] Mamabolo, M. A. and Myres, K. A detailed guide on converting qualitative data into quantitative entrepreneurial skills survey instrument. *The Electronic Journal of Business Research Methods*, pages 102–117, 2019.

[20] Okey, O. D., Melgarejo, D. C., Saadi, M., Rosa, R. L., Kleinschmidt, J. H., and Rodríguez, D. Z. Transfer learning approach to ids on cloud iot devices using optimized cnn. *IEEE Access*, 11:1023–1038, 2023.

[21] PINTO, G. E., Rosa, R. L., and Rodriguez, D. Z. Applications for 5g networks. *INFOCOMP Journal of Computer Science*, 20(1), 2021.

[22] Rastogi, R., Mondal, P., Agarwal, K., Gupta, R., and Jain, S. GA based clustering of mixed data type of attributes (numeric, categorical, ordinal, binary and ratio-scaled). *BVICA M's Int. J. Information Technology*, 7(2):861, 2015.

[23] Rich, T. S. South korean perceptions of unification: Evidence from an experimental survey. *Geo. J. Int'l Aff.*, 20:142, 2019.

[24] Rodriguez, D. Z., de Oliveira, F. M., Nunes, P. H., and de Morais, R. M. A. Wearable devices: Concepts and applications. *INFOCOMP Journal of Computer Science*, 18(2), 2019.

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

[26] Rosa, R. L., De Silva, M. J., Silva, D. H., Ayub, M. S., Carrillo, D., Nardelli, P. H., and Rodriguez, D. Z. Event detection system based on user behavior changes in online social networks: Case of the covid-19 pandemic. *Ieee Access*, 8:158806–158825, 2020.

[27] Rosa, R. L., Rodriguez, D. Z., and Bressan, G. Sentimeter-br: Facebook and twitter analysis tool to discover consumersâ sentiment. *AICT 2013*, page 72, 2013.

[28] 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 Trans. Industrial Informatics*, 15(4):2124–2135, 2018.

[29] Sadh, R. and Kumar, R. Clustering of quantitative survey data based on marking patterns. *INFOCOMP Journal of Computer Science*, 19(2):109–119, 2020.

[30] Sharma, U. and Manchanda, N. Predicting and improving entrepreneurial competency in university students using machine learning algorithms. In *Proc. 10th Int. Conf. Cloud Computing, Data Science & Engineering (Confluence)*, pages 305–309. IEEE, 2020.

[31] Silva, D. H., Rosa, R. L., and Rodriguez, D. Z. Sentimental analysis of soccer games messages from social networks using userâs profiles. *INFOCOMP Journal of Computer Science*, 19(1), 2020.

[32] Teodoro, A. A., Gomes, O. S., Saadi, M., Silva, B. A., Rosa, R. L., and Rodríguez, D. Z. An fpga-based performance evaluation of artificial neural network architecture algorithm for iot. *Wireless Personal Communications*, pages 1–32, 2021.

[33] Teodoro, A. A., Silva, D. H., Rosa, R. L., Saadi, M., Wuttisittikulkij, L., Mumtaz, R. A., and Rodriguez, D. Z. A skin cancer classification approach using gan and roi-based attention mechanism. *Journal of Signal Processing Systems*, 95(2-3):211–224, 2023.

[34] Velleman, P. F. and Wilkinson, L. Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician*, 47(1):65–72, 1993.

[35] Vichi, M., Cavicchia, C., and Groenen, P. J. Hierarchical means clustering. *Journal of Classification*, pages 1–25, 2022.

[36] 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.

[37] Zhang, Y. and Cheung, Y.-m. Learnable weighting of intra-attribute distances for categorical data clustering with nominal and ordinal attributes. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 2021.