Computer Science • 25(1) 2024 https://doi.org/10.7494/csci.2024.25.1.5685

Bhupendera Kumar
Rajeev Kumar

GENERALIZING CLUSTERING INFERENCES WITH ML AUGMENTATION OF ORDINAL SURVEY DATA

Abstract *In this paper, we attempt to generalize the ability to achieve quality inferences of survey data for a larger population through data augmentation and unification. Data augmentation techniques have proven effective in enhancing models' performance by expanding the dataset's size. We employ ML data augmentation, unification, and clustering techniques. First, we augment the limited survey data size using data augmentation technique(s). Second, we carry out data unification, followed by clustering for inferencing. We took two benchmark survey datasets to demonstrate the effectiveness of augmentation and unification. The first dataset contains information on aspiring student entrepreneurs' characteristics, while the second dataset comprises survey data related to breast cancer. We compare the inferences drawn from the original survey data with those derived from the transformed data using the proposed scheme. The results of this study indicate that the machine learning approach, data augmentation with the unification of data followed by clustering, can be beneficial for generalizing the inferences drawn from the survey data.*

Keywords survey research, ordinal data, data augmentation, clustering, unification, generalization

Citation Computer Science 25(1) 2024: 63–93

1. Introduction

Surveys are the most popular form of data collection in organizational and behavioral research [5]. The areas of policy-making, higher education, health care, psychology, and market research are some of the ones that commonly use surveys [15]. Correctly processing survey data has become a major problem due to the vast range of applications. Minor survey data analysis may occasionally produce bizarre results. Therefore, the right analytical tools are necessary to derive relevant insights from survey data. However, the nature of the data and the application's goal significantly impact how reliable analysis tools are [38]. It attempts to comprehend a phenomenon by compiling feedback from a sizable population [5].

The standard methods for analysis in survey research are statistical modeling tools for finding survey error(s); this necessitates prior knowledge of the association between the outcomes and covariates [46]. Unfortunately, in complex real-world circumstances where these interactions may not be accessible, it is not always possible to satisfy the condition of understanding the relationship mapping between the outcomes and variables. More adaptable modeling strategies are necessary for these situations that do not call for relational mappings to be predefined. Complex circumstances can be better understood by building relational mappings based on the inherent properties of the data [24]. For example, grouping data points according to their natural proximity can help us better comprehend a phenomenon, like the behavior of a sampled population. Flexible modeling techniques must be used, and numerous data-related issues must be resolved to extract relevant and trustworthy insights from survey data. Unique qualities of survey data include variability, hierarchical linkages, and the importance of category names [43]. Depending on the degree of heterogeneity, the data may contain a variety of metrics, including binary, continuous, categorical, or their mixtures.

Most survey techniques involve using a single mode of data collection. In today's complex world, single-mode survey techniques may not be sufficient. For example, universities survey students to learn about their perspectives, interests, and behavior to better understand the factors that contribute most to their entrepreneurial aptitude; the survey data could be multi-modal.

To address this, researchers employ multiple surveys allowing diverse inferencing and catering to complex themes. These surveys offer a range of methods, including mathematical analysis and qualitative inference, to gather comprehensive data and insights that align with the research objectives and complexities of the survey topic [6]. Unification is a process of combining various data elements to create an arrangement that is logical and consistent enough to allow for the drawing of reliable inferences. Since it makes it feasible to combine and bring diverse pieces of knowledge into one coherent whole, unification is crucial for effective inferencing. It must include patterns or connections into a single structure. Furthermore, unification calls for considering relevant factors affecting the discovered patterns or correlations [42]. The success of unification is crucial for accurate inference.

In addition, sampling is always limited by size, yet it is expected that such limited sampling should lead to the views of the whole population [33]. Participants might differ in their perspectives. Ensuring the sample size is enough to include all relevant viewpoints while performing qualitative research. A limited sample size can have a better chance of finding a wide range of impressions and increasing the credibility of their inferences; there could be fewer conflicts. However, analyzing such a range of data presents formidable difficulties [34].

Additionally, survey data generated by web-based survey software frequently contains small ordinal measurements. According to the research, treating values on small ordinal scales as value-based is improper [47]. On the other hand, methodologies that make use of vectors with ordinal values typically outperform pattern-based analysis techniques [40, 41]. Such techniques, widely used by most survey inferencing tools and techniques, lead to arbitrary inferencing. Apart from the facts above about the reliable analysis of survey data, the flexible modeling techniques, and the use of vector techniques, the survey faces the challenge of gathering information from the intended audience. It is one of the main problems with online surveys. Online surveys frequently require greater response rates, which could result in sufficient sample size and skewed findings [2].

In this work, we use machine learning (ML) techniques, namely, data augmentation, to augment the survey size. ML is mostly data-driven [30]. To put it another way, it offers adaptable modeling methods that exclusively rely on the intrinsic properties of the data to make the connections between the data and the results. ML usage may open survey research to generalized predictive modeling, limited to determining population features from a sample of data [8, 9]. Data augmentation is a generalization technique that enriches and enhances the population size for proper inference. We expect that the inferencing of the limited survey should be that of the population [22].

In this work, we focus only on ordinal data. But, like in many surveys, there is also associated numerical type data. So, as a result, we apply the unification process, which appropriately converts numerical values to ordinal values. After this, we put the dataset into a machine-learning model, especially for clustering. By doing this, we got better clusters for finding the effecting features from each cluster. It means the formed clusters are effective. In the result section, we show the efficiency measures of the clustering and can find suitable and effective features.

Therefore, we address the following research questions (RQs) in this work:

RQ1: Does the limited survey sampling be extended through ML augmentation reflecting a more significant population's general opinion?

RQ2: Does the augmented data with unification and clustering yield proper inferences?

Regarding the RQs mentioned above, the research in this work examines the proper inferencing obtained through augmentation and unification. For this purpose, we took two case studies, one for finding the competency factors in university

entrepreneurs and the second for breast cancer prediction features. The research contributions of this work are summarized as:

- We employ data augmentation techniques on *limited* survey to yield the inferencing of larger population.
- We identify the generalized driving features to determine the significant factors for prediction.
- The proposed ML methods yield significant inferences for ordinal-type survey data for proper decision-making.

The paper is organized as follows: Section 2 describes the motivation behind developing ML-driven methods. Section 3 briefly reviews the literature highlighting the data augmentation techniques and unification's role in clustering survey data. In Section 4, we describe, in detail, the concept of our proposed methodology. The experimental setup with dataset description of survey data and the detailed corresponding results are presented in Section 5.1. Finally, we conclude the paper in Section 6. Table 1 lists the key abbreviations that comprise this paper.

Table 1
Abbreviations

Abbreviation	Description
ADASYN	ADAptive SYNthetic
CNFL	Categorical to Numerical Feature Learning
DAUG	Data AUGmentation
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DNA	Deoxyribo Nucleic Acid
EM	Expectation Maximization
GA	Genetic Algorithm
GAN	Generative Adversarial Network
PCA	Principal Component Analysis
ROC	Receiver Operating Characteristics
SMOTE	Synthetic Minority Over Sampling
SOM	Self Organising Map
UFDM	Unification For Data Modelling

2. Motivation

The survey of a limited population should reflect the opinion of a large population. The survey aims to gather perceptions and viewpoints that may be applied to a more significant population. A properly chosen sample population that reflects the large population in terms of the pertinent features must be used to do this. The techniques may be used to draw valid conclusions about the attitudes and actions of a larger population. We are using ML techniques. It should be generalized. Therefore, we use the augmentation technique, which is a generalization technique. Even then, if we get

fewer responses, it may give loosely prominent results. So we have to expand our number of responses through augmentation. We depict an example of this process. Here, we assume only three features (A1, A2, and A3) and responses from two categories (A and B). We suppose that a survey is conducted on a questionnaire, and based on this, there are three features to observe (A1, A2, and A3). The features will be selected from three for decision-making on the two categories of responses (A and B).

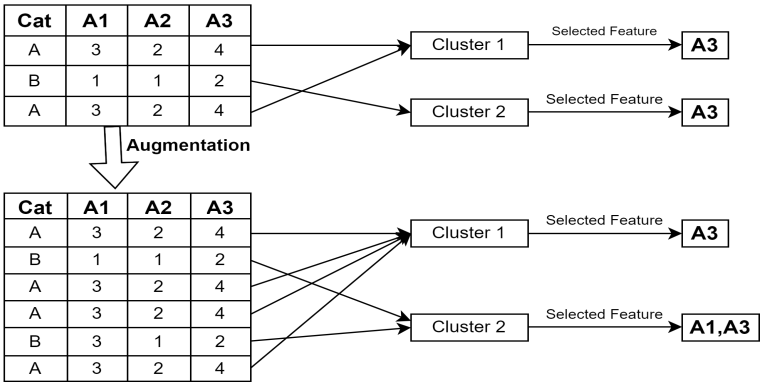

Figure 1. A Sample Example of Selected Features from a Survey Dataset

Figure 1 illustrates that in a small population size, before augmentation, the most governing feature for inferences through clustering is only A3. Considering the responses belong to two categories, A and B, they group into two clusters. A3 is most likely selected for inferencing from each cluster with maximum grading. On the other hand, after augmentation, the responses increased in number and were also grouped into two clusters. But this time, we got another feature A1 from cluster 2. There may be a possibility of getting all three features; this will be discussed in detail in the results section for limiting the selected features. Thus, we can get more generalized and precise inferences. Therefore, the inferences of ML techniques will be better suitable for this work.

3. Related work

3.1. Data augmentation in survey data

The fundamental objective of data augmentation is to create a productive and repeatable sampling method by adding concealed or unseen factors to the model. This technique gained prominence primarily in deterministic algorithms that aim to maximize likelihood functions or posterior densities with the expectation-maximization (EM) algorithm [16]. Constructing a data augmentation algorithm is somewhat of an art because data augmentation algorithms must be carefully developed for each model type [17]. Schliep and Hoeting introduced parameter-expanded data augmentation techniques to model ordinal data with the *probit* model. Specifically, the study

focused on implementing these algorithms for the probit linear mixed model in the context of spatially correlated ordinal response data. The researchers then demonstrated the applicability of the model by utilizing it to assess the biotic integrity of wetlands in Colorado [42].

Machine learning algorithms are typically evaluated based on their predictive accuracy. However, this approach may be unsuitable for imbalanced datasets where classes are not evenly represented or when the cost of different errors varies significantly. For instance, fraud detection often involves a class imbalance of 100 to 1, while other applications may have an imbalance of up to 1,00,000 to 1. Over-sampling techniques have been proposed to address this issue to balance the data. One approach involves creating synthetic examples of the minority class rather than simply over-sampling with replacement. This technique has been successful in handwritten character recognition, where operations like rotation and skew were used to perturb the training data and create additional examples. By generating synthetic examples, we can improve the training of machine learning algorithms on imbalanced data and ensure that the minority class is not overlooked. This approach can be precious in applications like fraud detection, where correctly identifying the minority class is critical. SMOTE (Synthetic Minority Over-sampling Technique) [12] demonstrates that a more effective classifier performance (in ROC space) can be achieved through a combination of our over-sampling method for the minority (abnormal) class and under-sampling for the majority (normal) class, compared to solely under-sampling the majority class.

The Adaptive Synthetic (ADASYN) [23] sampling approach has been developed to address these issues. The main idea behind ADASYN is to use a weighted distribution for different minority class examples based on their level of difficulty in learning. This means that more synthetic data is generated for minority class examples that are harder to learn than those that are easier to learn. As a result, ADASYN improves learning by reducing the bias introduced by class imbalance and adaptively shifting the classification decision boundary towards the difficult examples. Simulation analyses on several machine learning data sets have demonstrated the effectiveness of this approach across five evaluation metrics. The ADASYN sampling approach has emerged as a promising solution to this challenge. By generating synthetic data for minority class examples based on their level of difficulty in learning, ADASYN helps reduce bias and adaptively shift the classification decision boundary towards difficult examples. This approach effectively improves learning outcomes across various machine learning data sets, making it a valuable tool for tackling imbalanced data sets in modern data mining applications.

Temraz and Keane proposed a data augmentation method that generates synthetic, counterfactual instances in the minority class. Unlike other oversampling techniques that interpolate values between instances, this method adaptively combines existing instances from the dataset using actual feature values. To generate synthetic instances, the paper deploys a case-based counterfactual method. Counterfactual

methods are developed to generate posthoc examples to explain the predictions of black-box ML models and provide algorithmic recourse for end-users trying to mitigate automated decisions [45]. Hulse *et al.* analyzed eleven learning algorithms on thirty-five real-world datasets to guide machine learning practitioners and suggest future research directions on building classifiers from imbalanced data. This study is unique as no other related work has analyzed class imbalance on such a wide scope [48].

The data augmentation field is vast, and it is especially used in the field of images. Image data augmentation involves creating new images from existing ones by making small adjustments, such as changing brightness, rotating the image, or shifting the subject horizontally or vertically. This technique effectively increases a dataset's size and improves a machine-learning model's robustness. When a model performs differently on training data versus testing data, it's called generalizability. Overfitting occurs when a model has poor generalizability due to being overly trained on the training data. Simple transformations like horizontal flipping, color space augmentations, and random cropping were the earliest demonstrations of the effectiveness of Data augmentation. These transformations address invariances that pose challenges to image recognition tasks. The efficiency of geometric and photometric (color space) conversions was examined in comparative research by Taylor and Nitschke [44]. We looked at geometric changes, including flipping, 0° to 360° rotations and cropping, as well as color space transformations like edge improvement, PCA, and color jittering (random color manipulation). Eight thousand four hundred twenty-one photos with a size of 256×256 from the Caltech101 dataset were used in the 4-fold cross-validation test of the augmentations.

Generative modeling, nicknamed Generative Adversarial Network (GAN), is a fascinating data augmentation method. Generative modeling is constructing artificial instances from a dataset while maintaining the original set's features. The highly intriguing and enormously well-liked generative modeling framework known as GANs results from the above-mentioned adversarial training ideas. GANs are a means to "unlock" more information from a dataset, according to Bowles *et al.* [7].

3.2. Unification in survey data

After augmentation, another perspective is the unification. The challenges of unification rather than its benefits, particularly concerning long-term economic growth and the practical aspects of societal and political integration. The extent to which the vocabulary and understanding of unification are unknown is still uncertain [37]. There are various types of categorical data, such as text data, DNA sequences, and Census Bureau data, that humans easily understand. Still, many classification systems, like support vector machines (SVM), require numerical data representations. Most learning techniques transform categorical data into binary values to handle this, which can result in high dimensionality and sparsity.

CNFL uses eigen-decomposition to convert the proximity matrix into a reduced space that can be used for classification or clustering. It first employs simple matching

to measure the closeness between instances [21]. Mamabolo and Myres provided two significant contributions. Firstly, it outlines a precise and reproducible 8-step process for questionnaire development utilizing qualitative research, which enhances the methodology for mixed-method designs. Secondly, the study creates a research tool for measuring the extent of entrepreneurial skills. Ultimately, the findings offer implications for research methodology, entrepreneurship scholarships, and practical applications [32, 49]. In data analysis, it is expected to ask meaningless questions.

Understanding data scaling can sometimes help us identify nonsense, but we must use proper logic. Giordan and Diana developed a new clustering technique that addresses two common cluster analysis issues: group size selection and scale invariance. The method employs a multinomial model, a cluster tree, and a pruning approach to group objects. Two types of pruning are examined using simulations [20]. When dealing with real-world problems, data may include numeric and categorical variables. While many regression algorithms work well with numeric variables, categorical variables require additional considerations. However, decision tree algorithms can estimate targets based on specified rules and handle categorical and numeric variables. Kim and Hong proposed a new hybrid model combining a decision tree with another regression algorithm to analyze mixed data. The algorithm was evaluated on twelve datasets and achieved better or comparable accuracy to other methods without significantly increasing computational complexity [25–27].

The decision tree algorithm can handle categorical and numerical variables by evaluating the target based on predefined rules. This feature is used to create a new hybrid model that combines a decision tree with a different regression technique to analyze mixed data. The GA algorithm optimizes the new cost function and produces accurate clustering results. We can evaluate whether a GA-based clustering algorithm suits high-dimensional data collections with mixed features [36]. A novel distance metric is proposed to preserve 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 is devised to estimate the distance between categories of an ordinal attribute, which utilizes 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 [50].

3.3. Other techniques in survey data

Inference from sample surveys has traditionally focused on functions such as averages and totals of the findings made for the population's participants. However, in scientific applications, the superpopulation parameters linked to a stochastic mechanism assumed to produce the population's observations are frequently of more interest than the finite-population parameters. Even with a modest sampling proportion of the final units, cluster sampling, and conventional design-based variance calculations can significantly underestimate super-population variability [22]. In many empirical applications, there is a chance that mistakes may be associated with clusters. Thus,

it is crucial to strive for accurate statistical inference. We must make sure that our inference takes this into account. Usually, using conventional cluster-robust variance estimators is simple, but things may get complicated occasionally. The two main challenges are dealing with a small number of clusters and figuring out how to define the clusters [11]. Therefore cluster inferencing becomes a more crucial part of survey data analysis. Mixed datasets are frequently subjected to clustering to identify patterns and collect related objects for additional examination. However, it might be not easy to directly apply mathematical operations, such as summing or averaging, to the feature values of these datasets, making clustering mixed data tricky [3].

4. The proposed methodology

Multiple data sources may have different attributes when survey results are gathered. They could be nominal, numeric, or ordinal. Data of all kinds affect survey research. Our conclusions will be more reliable and useful if we incorporate all available facts. One more aspect is there while collecting the data. The number of responses may be small compared to getting a better result with more respondents. In this section, we proposed a model to conquer these deficiencies. The proposed design workflow of this model is given in Figure 2. In the following Subsections, we describe each process of this workflow.

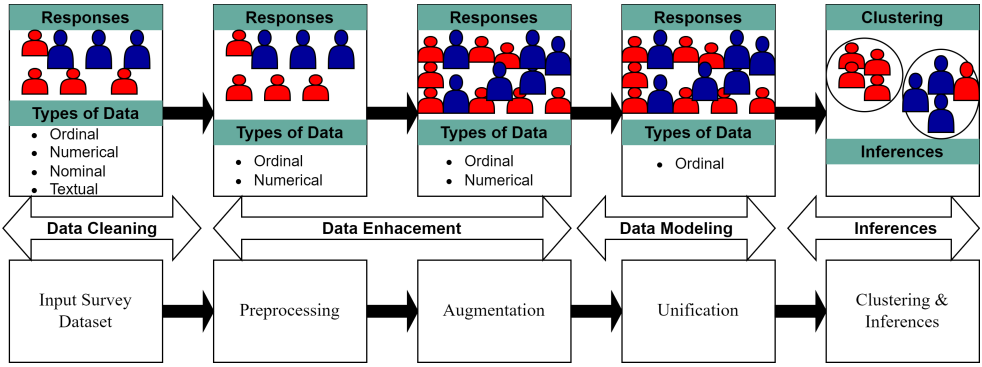

Figure 2. Workflow of the proposed methodology

4.1. Preprocessing

Survey data is essential for preparing the dataset for in-depth analysis and modeling. The reliability and validity of research findings may be increased by resolving difficulties and conflicts for improved data quality, standardization, and representativeness resulting in insightful findings that support well-informed decision-making.

Preprocessing survey data is a vital and complex phase that aims to ensure the gathered data is precise, consistent, and prepared for insightful analysis. Data cleansing is when possible mistakes and missing values are found and dealt with

properly. If attribute values are lacking during the process, the median value based on domain knowledge will fill any gaps. In this article, we focus on two different categories of data, numerical and ordinal, among many others. So the first preprocessing step in this scenario is taking the ordinal and numerical data from the surveyed dataset. After preprocessing and cleaning, we get the dataset for further use. We named it the original dataset.

4.2. Augmentation

We have mentioned SMOTE and ADASYN in the related work section. These two are well-known techniques for data augmentation. These techniques use the nearest neighbor for class imbalance problems with at least two classes. In this article, we do not deal having class imbalance problems. So in this part, another augmentation, a machine learning approach, is used to add more statistical techniques to the already-existing data to expand the diversity of the data. This enhances the generalization and effectiveness of the model. Through augmentation, we attempt to achieve the quality of survey data for a larger population. The Data AUGmentation (DAUG) Algorithm (Algorithm 1) is the pseudo-code for augmentation.

Algorithm 1 *DAUG (S,m,n,P)*

Input: Dataset S, Number of rows m, Number of Columns for using deviation l
Output: Augmented Dataset P

dataset $S[\,]$: Select the numerical and ordinal attributes
n : size of S
m : Number of rows randomly selected from n for augmentation
if $m > n$ **then**
    Reduce the size of m
end if
$m_1[\,]$: make a sample copy of m rows
$m_2[\,]$: make a sample of m rows with random ordinal data distribution
m : Concatenate $m, m_1 \& m_2$
l : Number of columns for considering deviation
for i to l **do**
    Column_medium[i] : choose the medium from each column
    deviation[i] : the deviation for the selected medium from the respected column in each column
    calculate average_deviation[i]
end for
k : Number of rows to add based on average deviation and median
$m = n \times k$
add m rows to dataset P
Normalize P
return P

By applying the DAUG (Algorithm 1), the survey dataset is expended. The original dataset is passed to the model. In the first step, the attributes with numerical and ordinal values are selected and treated as the original dataset. Select the number of rows randomly from the original dataset suitably. Then make two copies of the selected raw data, one for replication and another for different data that have changed ordinal values. Then combine these copies for the augmentation process based on the row-wise mean and standard deviation.

4.3. Unification for data modeling (UFDM)

After augmentation, we consider two types of data values, numerical and ordinal. We make an effort to incorporate survey data that is numerical and ordinal. We use a Gaussian distribution to represent the data. Therefore, we first transform the numerical data into ordinal data that follows the distribution. This process is called a unification for data modeling. The UFDM for unification is given in Algorithm 2.

Algorithm 2 *UFDM (S, a_n)*

Input: Dataset S, Numerical attribute a_n
Output: Unified Dataset D

Select a_n from S
min_a_n : Minimum of the numerical attribute
max_a_n : Maximum of the numerical attribute
avg_a_n : Average of the numerical attribute
$temp\_count[\,]$: for number of occurrence of each number
for i to each number in range min_a_n to max_a_n **do**
    $temp\_count[\,i]$
end for
for i to each row in a_n **do**
    num := row.num
    $temp\_count[\,i] := temp\_count[\,i] + 1$
end for
if $temp\_count[\,]$ is left skewed **then**
    Assignment of ordinal values with making bin following the increasing bin-size from left to right
else if $temp\_count[\,]$ is right skewed **then**
    Assignment of ordinal values with making bin following the decreasing bin size from left to right
else
    Assignment of ordinal values with making bin following the equal bin size from left to right
end if
Replace numerical values with ordinal values and update the dataset with named D
return D

Applying the algorithm UFDM (Algorithm 2), the numerical values are converted to ordinal values through the unification process in the model. The generated dataset from the augmentation process is the input for the unification process. Find the statistics of this dataset, like the minimum, maximum, and average of each numerical attribute. We want to convert numerical data to ordinal data. Then find and count the number of occurrences of each element in ascending order of minimum to maximum values. Adjust the bin size of the ordinal valued bins (based on the Likert scale) accordingly for the skewness nature of the dataset.

4.4. Clustering and inferencing

After augmentation and unification for data modeling, we compare the efficiency of groups through clustering. Clustering entails grouping instances into clusters based on similarity to discover underlying patterns or structures within the dataset. K-means algorithm seeks to optimize the cluster allocations by minimizing the sum of squared distances between data points and their associated centroids. So we use K-means clustering for the whole process. We apply K-means at three levels at the original dataset, after augmentation, and after unification. The clusters made during the process should have improved quality for inferencing so that generalized features can be stated. The governing generalized feature selection is the main focus of inferencing.

4.5. Complexity analysis

The workflow of the proposed technique encapsulates three techniques, namely, Augmentation, Unification, and Clustering. In this subsection, we estimate the computational complexity of the proposed methodology. Let n be the number of rows in the original dataset S.

1. **Augmentation (DAUG):** The steps of the augmentation algorithm (DAUG) are listed in Algorithm 1. The standard deviation in the associated columns is taken into consideration for selecting rows for augmentation Algorithm lines 6–14 are used to calculate each attribute's computation. The time complexity for selecting m rows for augmentation from the dataset S is $O(m)$. The number of columns for considering deviation is l. These columns with selected rows are augmented, therefore, the complexity for this process is $O(m * l)$. The last step is appending the number of k rows, the complexity is $O(k)$. Therefore, the complexity of DAUG Algorithm (Algorithm 1) is $O(m + m * l + k) \approx O(n^2)$.
2. **Unification (UFDM):** The next stage is the unification work: the UFDM Algorithm 2. In UFDM, lines 2–12 are for the unification, and lines 13–20 are for the assignment. Let a_n be the number of numerical attributes. The range of numerical values is in r. The complexity for finding the minimum, maximum, and average of each attribute is $O(a_n * n)$. The complexity for fitting the ordinal values according to the range of numerical values is $O(r * n)$. The last step to replacing the numerical values with corresponding ordinal values

is $O(n * 1)$ complexity.Therefore, the complexity of UFDM Algorithm (Algorithm 2) is $O(a_n + r * n + n) \approx O(n^2)$.

3. **Clustering:** Let the number of desired clusters be (t), the number of rows to be clustered be (n), and the number of iterations until convergence be given by (i). The number of the attributes (a_n) determines the complexity of the K-means method. The Clustering is of $O(t * n * i * a_n) \approx O(n^3)$ [1, 29, 35, 51].

K-means is susceptible to the presence of outliers and is known to perform poorly in the presence of outliers. However, there are several other clustering algorithms, e.g., DBSCAN, hierarchical clustering [18], etc. that handle outliers at the cost of higher complexity [29]. However, this is the future direction of this work.

5. Experimental results and analysis

In this section, we experiment with two datasets and use them to illustrate the proposed methodology and select the generalized features. The performance of clustering algorithms can be assessed using a wide range of metrics, which are utilized depending on a particular task and objectives of the clustering method. In this work, we have considered three performance metric measures: the Silhouette scores [39], Calinski Harabasz Index [10], and Silhouette Analysis plot [39].

Silhouette scores. The silhouette score calculates how well each data point fits into its allocated cluster. This is calculated as the ratio of the mean distance between a data point and all the remaining data points in a comparable cluster to the average distance between a data point and all similar data points in the closest cluster. A higher silhouette score means that the data points have been successfully divided into different clusters that are uniform inside and well-separated by the clustering method.

Calinski Harabasz index. On the contrary, a higher score on the Calinski-Harabasz index denotes superior clustering efficiency. It evaluates the ratio of around-cluster variation to within-cluster variance, which implies how well the Calinski-Harabasz index consider both the gap between clusters and the compactness of each cluster.

Silhouette analysis plot. Each data point's silhouette scores are displayed on the silhouette analysis plot, showing the way each one fits into the cluster to which it was assigned. The range of a silhouette score is from -1 to 1: A clustering allocation with an average of $+1$ is considered successful, whereas one with a value of 0 is considered unclear. A good clustering solution has most data points near $+1$, denoting clearly defined clusters, whereas a not-good clustering solution has values close to 0 or negative values, signifying overlaps or incorrect assignments. By finding the clusters with the greatest average silhouette score representing the most distinct and well-separated, the plot aids in determining the ideal number of clusters. It sheds light on how the quality of clustering and cluster numbers are traded off.

Experimental setup. We have used the proposed model and the clustering technique in the Anaconda edition of Python 3.7 on Windows 10 PC with an Intel Core $i5$ CPU ($2.0GHz$) and $4GB$ of RAM and 64-bit operating system, x64-based processor. In addition to *sklearn*, *metplotlib*, the *pandas* are also used for reading data and visualizing it graphically. We enhanced a Python module of our model to allow for simple code implication. The experiment was conducted within Jupyter Notebooks, using its open-source libraries to speed up and simplify the development process.

Datasets. For our experiment, we have taken two benchmark datasets that are freely available. These datasets are collected from the surveys. These datasets can be downloaded from *Kaggle*, a website with modeling and analysis competitions where data miners compete to create the most effective models using data posted by businesses, researchers, and other users. The following datasets are taken:

- Dataset I: Entrepreneurial Competency Survey.
- Dataset II: Breast Cancer Survey.

5.1. Dataset I: entrepreneurial competency survey

We have collected a dataset [28] to accomplish insightful information about the connection between university students' entrepreneurial habits. This survey aimed to gather data for the students' entrepreneurial propensities levels. Two hundred nineteen responses from survey respondents who were university students make up the dataset we used for this study. Different abbreviations are used for the dataset. These are briefly listed in Table 2.

Table 2
Abbreviation used for Education Sector and Features

Education Sector	Abbr.	Features	Abbr.
Art, Music or Design	AMD	Age	A1
Economic Sciences, Business Studies, Commerce and Law	ESBSCL	Perseverance	A2
Engineering Sciences	EC	DesireToTakeInitiative	A3
Humanities and Social Sciences	HSS	Competitiveness	A4
Language and Cultural Studies	LCS	SelfReliance	A5
Mathematics or Natural Sciences	MNS	StrongNeedToAchieve	A6
Medicine, Health Sciences	MHS	SelfConfidence	A7
Others	OT	GoodPhysicalHealth	A8
Teaching Degree (e.g., B.Ed)	TD		

5.1.1. Dataset description

This dataset, which has two hundred nineteen instances, comprises nine features in the form of attributes, i.e., Age (A1), Perseverance (A2), DesireToTakeInitiative (A3), Competitiveness (A4), SelfReliance (A4), StrongNeedToAchieve (A6),

SelfConfidence (A7), GoodPhysicalHealth (A8), and EducationSector. Age is the numerical data. Perseverance, DesireToTakeInitiative, Competitiveness, SelfReliance, StrongNeedToAchieve, SelfConfidence, and GoodPhysicalHealth are in ordinal data. EducationSector is categorical data.

5.1.2. Statistical analysis

All these features and their overall and attribute-wise mean and standard deviations received from the survey are given in Table 3.

Table 3
Overall and Attribute-wise mean and standard deviations of Original Survey Data

Education Sector		Attributes							
		A1	A2	A3	A4	A5	A6	A7	A8
AMD	Mean	20.33	3.19	3.38	3.43	3.57	3.76	3.67	3.38
	StdDev	1.21	1.01	1.40	1.22	1.14	1.23	1.17	1.25
ESBSCL	Mean	19.56	3.38	3.72	3.47	3.75	4.09	3.56	3.63
	StdDev	1.64	0.96	1.04	1.09	0.94	1.04	1.09	1.32
ES	Mean	19.74	3.38	3.72	3.72	3.81	4.02	3.62	3.61
	StdDev	1.23	1.01	1.02	1.02	0.98	0.90	1.10	0.99
HSS	Mean	19.60	3.40	3.60	3.00	4.00	3.80	3.60	3.60
	StdDev	0.80	0.80	1.02	1.41	1.10	0.98	1.02	1.02
LCS	Mean	19.00	3.00	5.00	3.00	3.00	5.00	5.00	2.00
	StdDev	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MNS	Mean	18.75	3.00	3.25	3.25	2.25	3.00	3.25	3.75
	StdDev	1.17	0.89	1.50	0.98	1.20	1.21	0.81	1.21
MHS	Mean	19.60	3.40	3.20	3.40	3.90	3.60	3.50	3.70
	StdDev	1.20	1.20	1.66	1.56	1.22	1.36	1.28	1.27
OT	Mean	20.00	3.25	3.35	3.45	3.45	3.35	3.30	3.30
	StdDev	0.95	0.83	1.28	1.02	1.07	0.96	1.05	0.95
TD	Mean	19.00	4.00	3.67	3.67	3.67	4.00	3.33	3.67
	StdDev	0.82	0.82	1.25	1.25	1.25	0.82	1.70	1.25
Overall	Mean	19.75	3.35	3.62	3.59	3.72	3.91	3.58	3.56
	StdDev	1.29	0.99	1.15	1.11	1.05	1.02	1.12	1.10

Survey data is augmented with the help of the proposed data augmentation techniques (Algorithm 1) to increase the size of the dataset. Table 4 shows the augmented dataset's overall and attribute-wise mean and standard deviations, with 1676 instances.

Table 4
Overall and Attribute-wise mean and standard deviations of Augmented Survey Data

Education Sector		Attributes							
		A1	A2	A3	A4	A5	A6	A7	**A8**
AMD	Mean	20.48	3.23	2.70	3.11	3.43	3.89	3.95	3.48
	StdDev	1.31	1.04	1.39	1.25	1.07	1.05	1.21	1.31
ESBSCL	Mean	19.53	3.27	3.01	3.05	3.56	4.07	3.93	3.55
	StdDev	1.55	0.90	1.37	1.10	0.87	0.91	1.14	1.29
ES	Mean	19.70	3.41	2.89	3.36	3.65	4.06	3.98	3.65
	StdDev	1.18	0.98	1.30	1.15	0.94	0.76	1.08	0.96
HSS	Mean	19.73	3.45	2.95	2.64	3.64	3.73	3.82	3.64
	StdDev	0.86	0.78	1.30	1.23	1.07	0.86	1.11	0.98
LCS	Mean	19.00	3.00	3.50	2.67	3.00	4.67	5.00	2.00
	StdDev	0.00	0.00	1.19	0.47	0.00	0.47	0.00	0.00
MNS	Mean	18.50	3.00	3.25	3.00	2.40	3.20	3.60	3.80
	StdDev	1.28	1.00	1.24	1.10	1.20	0.98	0.80	0.98
MHS	Mean	19.50	3.44	3.11	3.06	3.78	3.56	3.67	3.72
	StdDev	1.38	1.12	1.33	1.47	1.08	1.17	1.20	1.15
OT	Mean	20.08	3.37	2.99	3.08	3.42	3.61	3.87	3.42
	StdDev	1.01	0.78	1.31	1.06	0.91	0.81	1.10	0.94
TD	Mean	18.71	4.00	2.71	3.14	3.43	4.00	3.71	3.57
	StdDev	0.70	0.93	1.46	1.36	1.18	0.76	1.75	1.40
Overall	Mean	19.74	3.37	2.92	3.22	3.56	3.96	3.94	3.59
	StdDev	1.28	0.96	1.32	1.18	0.99	0.88	1.12	1.07

Error bar line graphs. Error bar line graphs are used to visualize and analyze data and provide essential insights into a dataset's consistency and variability. The distribution of the data around the mean value is revealed by these graphical representations, which aid in determining the relevance of the gathered data. The size of the error bar line graphs, which are frequently represented by standard deviation, effectively conveys how far a given data point deviates from the mean. A small standard deviation bar denotes minimal variability and a higher degree of confidence in the correctness of the data. It also indicates that the data points are closely grouped around the mean. On the other hand, a bigger standard deviation bar highlights greater variability and maybe more uncertainty by showing a wider range of data points away from the mean.

The comparison for the original and augmented data error bar line graphs is shown in Figure 3. This shows the attribute-wise comparison. The blue lines are for the original dataset, and the red lines are for the augmented dataset. In most attributes, the overlapping area shows that the augmented dataset does not deviate from the original data. This process can access the augmented dataset as a large population. And the inferences from the augmentation process we get are the more generalized inferences to make decisions.

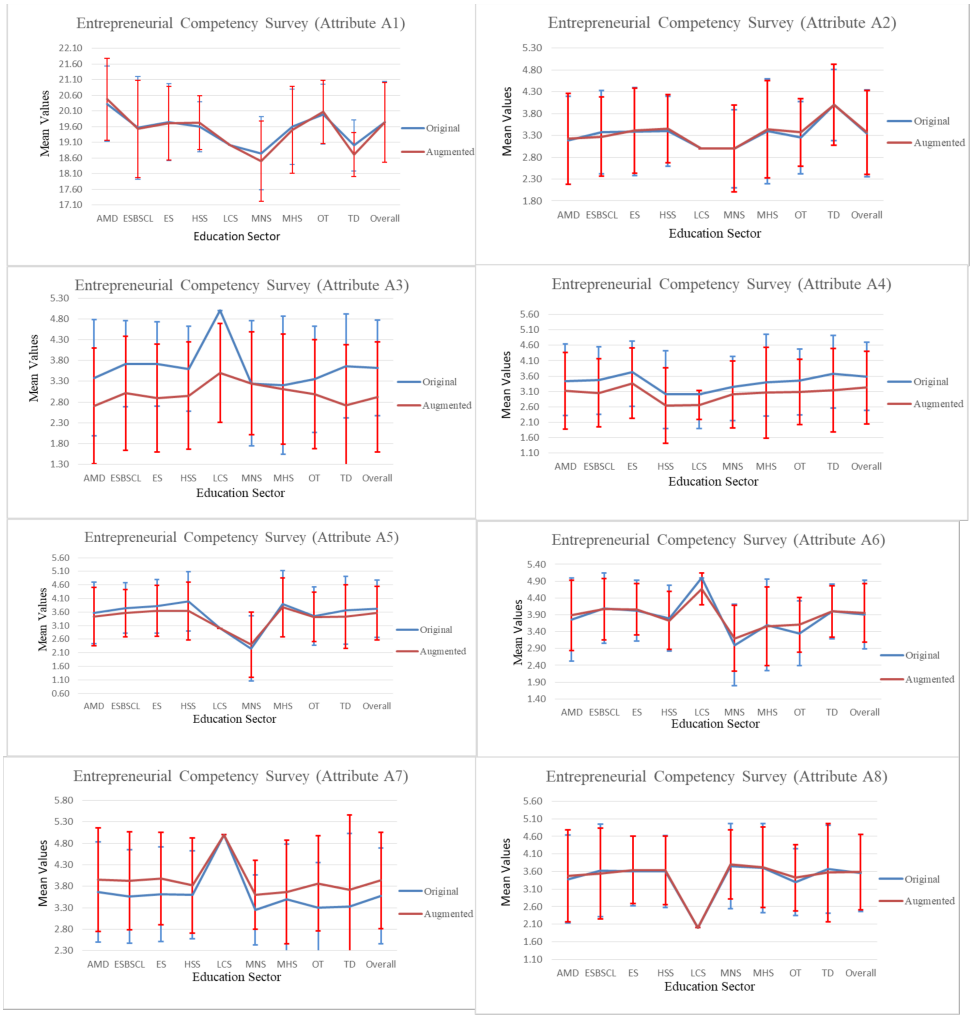

Figure 3. Error bar line graphs for Entrepreneurial Competency survey

5.1.3. Clustering results

We explore our experiment in three parts. In the first part, we considered the original dataset. First, we preprocessed the dataset and separated numerical and ordinal type attributes from the dataset. Then we apply the K-means algorithm for two to nine clusters on the original dataset collected from the survey.

As we have only two hundred nineteen instances in the dataset. So in the second part, we apply the data augmentation techniques to have a proper number of instances. In this manner, we will have sufficient instances and expect to get better inferences through clustering. In the third part, we convert the numerical data to the ordinal data according to Gaussian distribution. After unification, we again apply the K-means algorithm to find the clustering behavior.

In Figures 4, 5, and 6, we have a Silhouette Analysis plot on original, augmented, and unified datasets, respectively. The measurements are shown on two to nine clusters using the K-means algorithm. The resulting plot displays the mean silhouette score for every single clustering solution and the silhouette scores for each cluster data point. High silhouette scores for every point of data and an elevated average silhouette score are desirable as they demonstrate that the data points are correctly segregated and clustered.

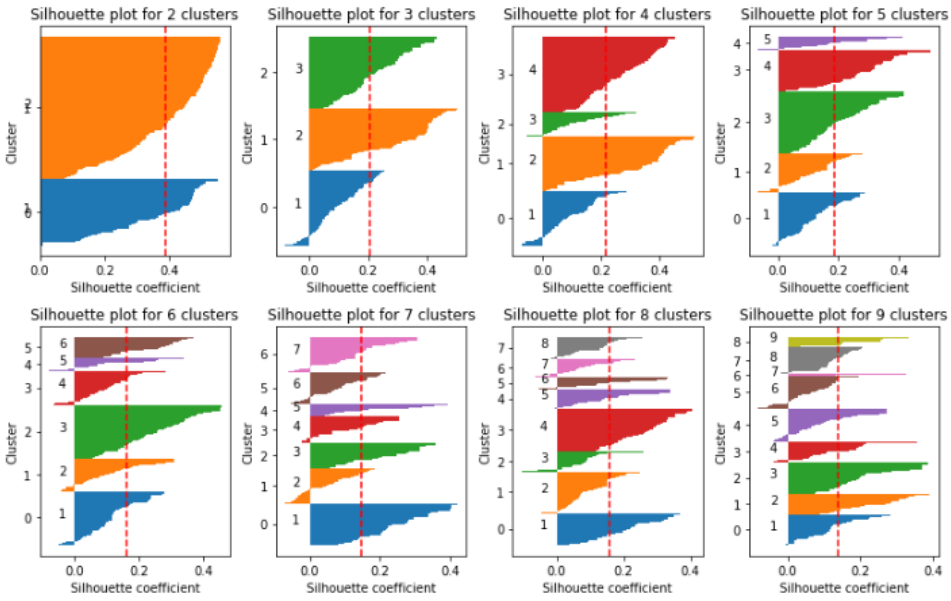

Figure 4. Silhouette analysis plots for original dataset clustering

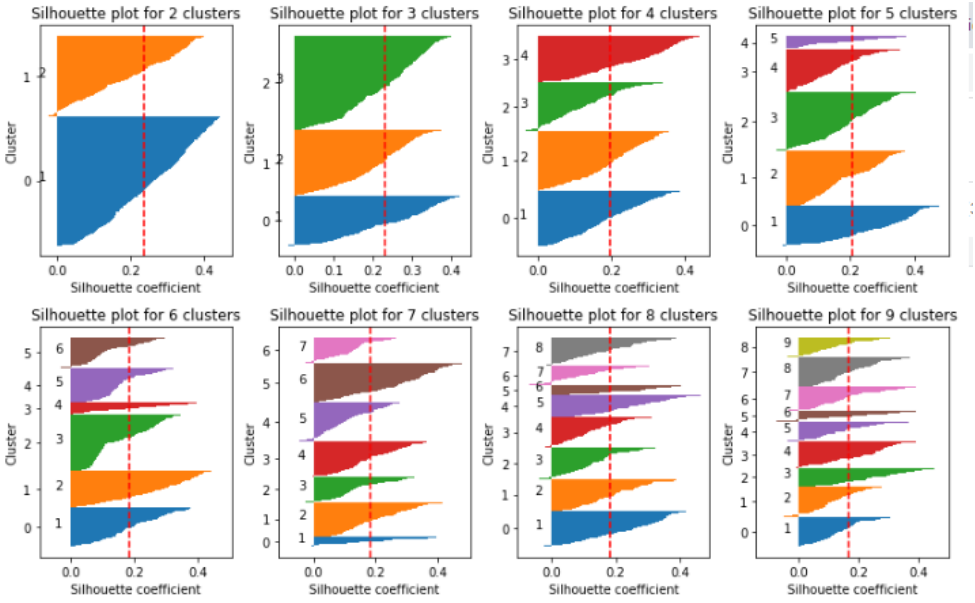

Figure 5. Silhouette analysis plots for augmented dataset clustering

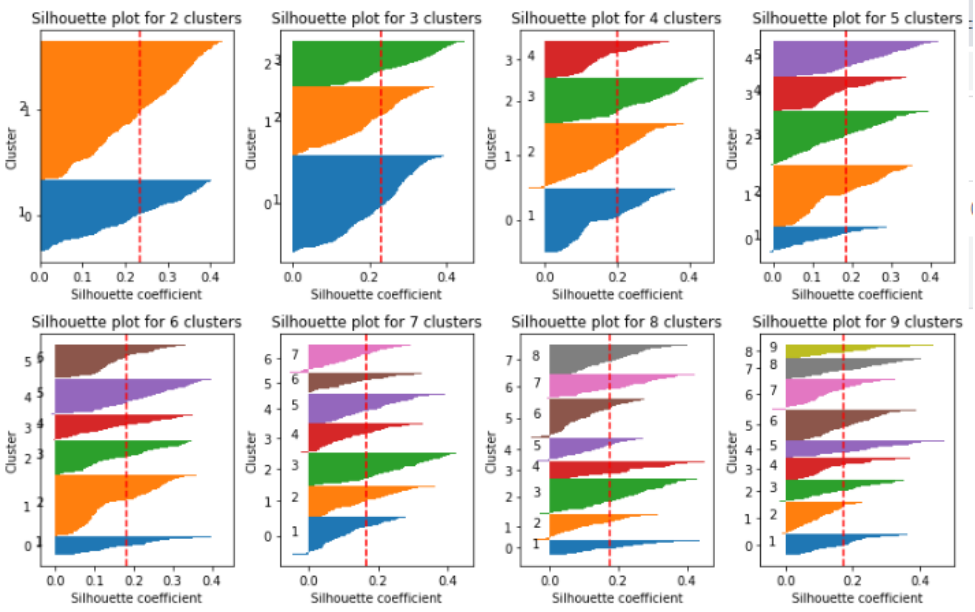

Figure 6. Silhouette analysis plots for unified dataset clustering

5.1.4. Result analysis

In Figure 7, the average Silhouette scores and Calinski-Harabasz indices are shown for the three datasets: one is the original, the second dataset is the one after the augmentation dataset, and the third one is the dataset after unification on three to nine clusters; these scores suggest that higher scores at the same number of the cluster have more data points that were successfully divided into different clusters which are similar inside and distinct from one another.

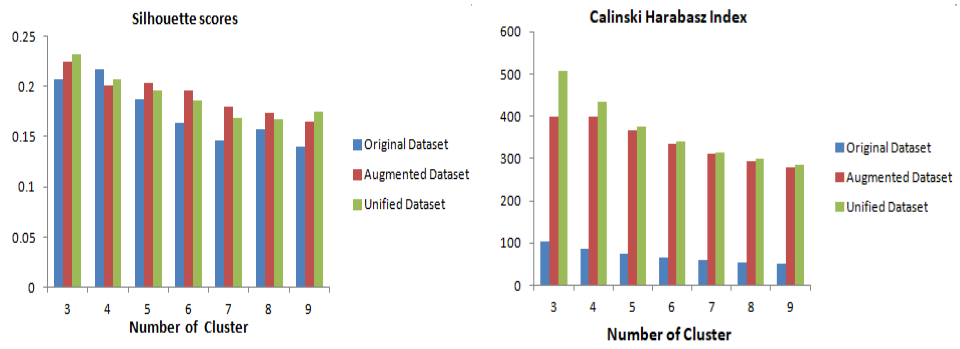

Figure 7. Performance metrics

For example, in Figure 7, at cluster nine, we have Silhouette scores of 0.1397 for the original dataset, 0.1651 for the Augmented dataset, and 0.1746 for the Unified dataset. This means efficiency is improved in clustering after augmentation and unification. The Calinski-Harabasz index considers both the distance between clusters and the compactness of each cluster. The index is higher at each cluster after, one by one, augmentation and unification. Figures 4, 5, and 6 are the complete measurement of Silhouette scores for each data point in each cluster, as well as the average silhouette score for the entire clustering of original, augmented, and unified dataset.

These outcomes show that the clustering procedure has successfully assigned each response to the cluster most closely resembling its features. It produces well-defined clusters with internally comparable replies and clear distinctions between other groups. So, the inferences from these datasets are shown in Table 5. We collected more generalized inferences by ML technique, augmentation.

Table 5
Selected Features for Entrepreneurial Competency

Dataset	Inferences (Competency Factors)
Original Dataset	StongNeedtoAchieve, Desireto TakeInitiative
Augmented Dataset	StongNeedtoAchieve, DesireTo TakeInitiative, Self Confidence
Unified Dataset	StongNeedtoAchieve, DesireTo TakeInitiative, Self Confidence

5.1.5. Runtime analysis

As shown in Figure 8, the state-of-the-art, SOM takes a longer time to train, particularly for big and high-dimensional datasets. It depends on the variables like the amount of data, the network, and the quantity of training iterations [4]. The size and density of the dataset affect how long DBSCAN takes to run. It may be less effective on large datasets, but it works well on datasets with different cluster densities [31]. That is why, we got higher runtime in our dataset available in Figure 8. Our methodology with K-means: among the three, the K-means is frequently the quickest. The convergence speed, which is determined by the start centroids and data distribution, might, however, affect the actual time.

We assess the runtime of the Entrepreneurial Competency dataset in the following Figure 8 for these three techniques at various cluster counts. We used the running time in seconds for SOM and our methodology while we used a logarithmic scale of time for DBSCAN techniques. It can be seen that the proposed technique performs better than the SOTA methods.

#Clust	SOTA		Proposed
	SOM	DBSCAN	
3	5.0	41.2	1.3
4	5.2	41.1	1.1
5	5.0	40.9	2.1
6	5.0	41.3	1.3
7	4.9	40.7	1.2
8	5.0	41.6	1.4
9	4.9	41.0	1.3

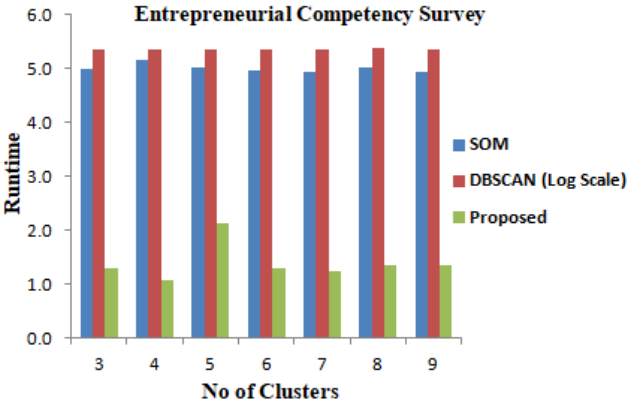

Figure 8. Runtime comparision for Entrepreneurial Competency survey

5.2. Dataset II: breast cancer survey

Next, we have taken a benchmark breast cancer survey dataset for this case study. In this dataset, there are six hundred ninety-nine responses. The dataset values are ordinal. We considered nine features for our study. The dataset, made up of clinical cases Dr. Wolberg documented, is distinguished by the data's arrival time. The dataset attempts to make it easier to forecast the occurrence of breast cancer. An individual code number that serves as an identification for each sample represents it. The features of each sample are then described using a set of Nine attributes. These characteristics include numerical measurements with a range of one to ten.

5.2.1. Dataset description

The survey aimed to gather data on Breast Cancer prediction. The dataset we used for this study comprises six hundred ninety-nine responses from survey respondents. Different abbreviations are used for the dataset. These are briefly described in Table 6. It is compassing attributes such as Clump Thickness (B1), Uniformity of Cell Size (B2), Uniformity of Cell Shape (B3), Marginal Adhesion (B4), Single Epithelial Cell Size (B5), Bare Nuclei (B6), Bland Chromatin (B7), Normal Nucleoli (B8), and Mitoses (B9). In combination, these characteristics capture crucial cell behavior and morphology features that point to probable malignancy. The construction and assessment of breast cancer prediction models are therefore made possible by the extensive set of features with associated diagnostic labels provided by this dataset.

Table 6
Abbreviation used for Breast Cancer survey

Features	Abbr.
Clump Thickness	B1
Uniformity of Cell Size	B2
Uniformity of Cell Shape	B3
Marginal Adhesion	B4
Single Epithelial Cell Size	B5
Bare Nuclei	B6
Bland Chromatin	B7
Normal Nucleoli	B8
Mitoses	B9

5.2.2. Statistical analysis

The mean values and standard deviations of these features in original and after ML techniques, augmented data are given in Table 7.

Table 7
Features and their corresponding mean values and standard deviation

Dataset		Attributes								
		B1	B2	B3	B4	B5	B6	B7	B8	B9
Original data	Mean	4.42	3.13	3.21	2.81	3.22	3.54	3.44	2.87	1.59
	StdDev	2.81	3.05	2.97	2.85	2.21	3.64	2.44	3.05	1.71
Augmented data	Mean	4.02	2.98	3.16	2.95	3.40	3.80	3.84	3.39	2.42
	StdDev	2.35	2.88	2.76	2.73	2.13	3.51	2.55	3.26	2.85

Error line bar graphs. Earlier, we mentioned the error line bar graphs in Subsection 5.1.2. Here we give some information on the Breast Cancer survey dataset. Figure 9 compares the original and augmented data error Line Bar Graphs. The comparison of attributes is also demonstrated here. The red lines represent the augmented dataset, and the blue lines represent the original dataset. The overlapping region for most attributes demonstrates that the enhanced dataset does not diverge from the

original data. This technique allows access to the enormous population of the expanded dataset. The inferences we draw from the augmentation process are those that are more broadly applicable to a significantly large population.

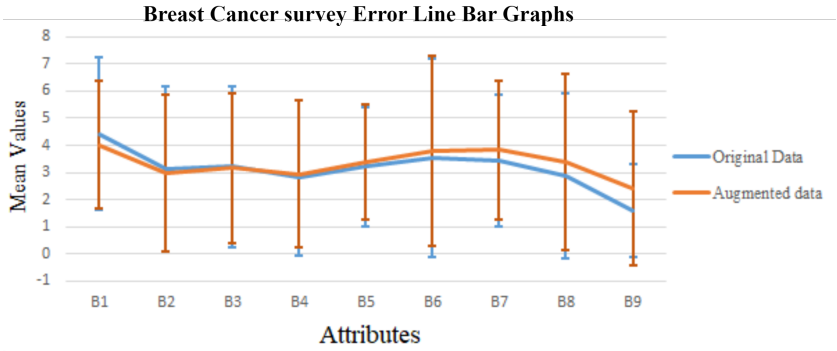

Figure 9. Error line bar graphs for Breast Cancer survey

5.2.3. Clustering results

We break up our investigation into two sections. We use the original dataset in the first section. The dataset was initially preprocessed and then characteristics of the ordinal type were extracted. The original dataset gathered from the survey is then subjected to the K-means method for two to nine clusters. Since the dataset has 699 occurrences only, we use the proposed data augmentation techniques (Algorithm 1) in the second section to have the right number of instances. In this way, we will have enough examples and may use clustering more effectively to draw more accurate conclusions. To learn more about the behavior of the clustering, we employ the K-means technique. The K-means technique displays the measurements on two to nine clusters. The resultant figures show the silhouette scores (see Fig. 10) for each cluster data point and the mean silhouette score for each clustering solution.

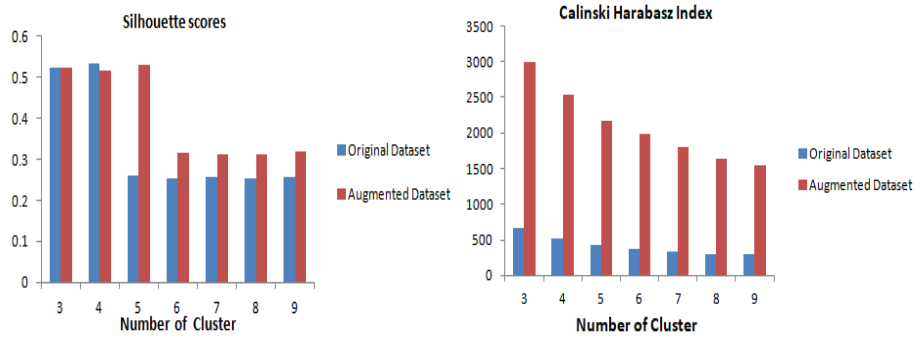

Figure 10. Performance metrics

Every data point should have a high silhouette score, and the average silhouette score should be high since these metrics show that the data points are appropriately grouped and clustered (see Fig. 11, 12).

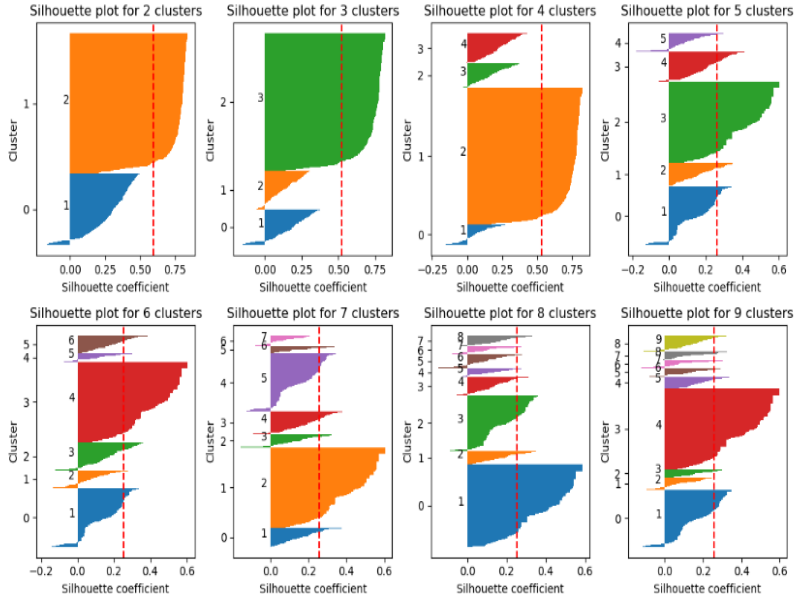

Figure 11. Silhouette analysis plots for clustering of the *original* dataset

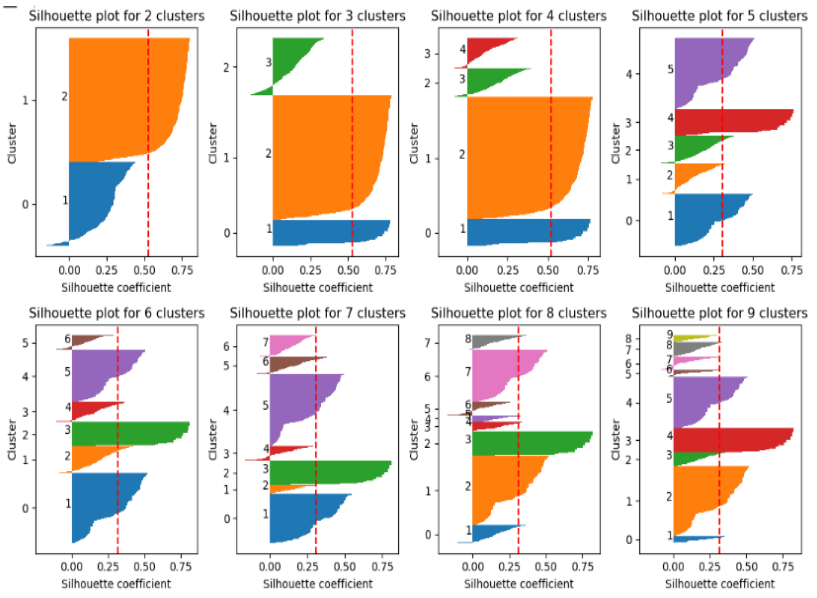

Figure 12. Silhouette analysis plots for clustering of the *augmented* dataset

5.2.4. Result analysis

The results demonstrate that each response was effectively allocated to the cluster that best matched its characteristics. This resulted in well-defined clusters with internally similar responses and noticeable differences between groups. Therefore, the conclusions drawn from these datasets are displayed in Table 8. We utilized the ML approach of augmentation to acquire more generalized inferences.

Table 8
Selected features for the Breast Cancer survey data

Dataset	Inferences (Selected Features)
Original Dataset	Bare Nuclei, Clump Thickness
Augmented Dataset	Bare Nuclei, Clump Thickness, Uniformity cell size, mitoses

5.2.5. Runtime analysis

We have mentioned the runtime analysis for the SOM and DBSCAN in Subsection 5.2.5. In the following Figure 13, we measure the runtime of the Breast Cancer Survey dataset for these three approaches at different cluster counts. For the SOM and the proposed methods, we take runtime in seconds; for DBSCAN techniques, we use the logarithmic scale for time. We may conclude that our strategy outperforms the SOTA techniques.

#Clust	SOTA		Proposed
	SOM	DBSCAN	
3	7.5	97.7	1.0
4	7.6	98.5	1.5
5	7.7	96.9	1.3
6	7.6	97.5	1.3
7	7.7	97.6	1.3
8	7.7	97.7	1.3
9	7.7	97.5	1.2

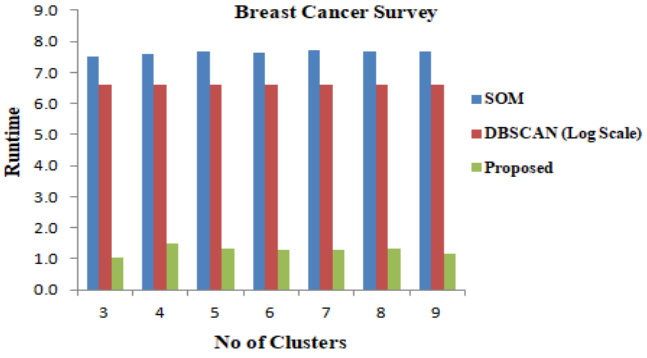

Figure 13. Runtime comparision for Breast Cancer survey

5.3. Discussion

The results obtained with the proposed method suggest that the quality and dependability of inferential findings for a large population from a small population can be improved using augmentation and unification procedures. By extending the existing data with methods such as mean and standard deviation, augmentation improves the dataset's representativeness. Augmentation lowers the chance of bias and improves the generalizability of the inferences made from the analysis. On the other side,

unification refers to combining numerical and ordinal datasets. Unification enables the fusion of many viewpoints and data modalities. Insightful findings, more reliable forecasts, and more precise modeling can result from this.

Scalability. We have proposed three algorithms in this paper. If we consider fewer attributes, the time complexity of the augmentation and unification algorithms is nearer to linear; however, while applying K-means, it may be quadratic complexity and needs to be addressed for scalability. On the other hand, DBSCAN is of quadratic complexity which may also need to be addressed [14]. Since DBSCAN, despite quadratic complexity is made scalable; the proposed algorithms lie in between linear to quadratic and could be made better scalable. This is an area of future work.

In Subsection 4.5, we have discussed computational complexity, which is polynomial between linear and quadratic for both, the augmentation and unification algorithms, and it is very obvious that as the scaling happens the time increases. So, it is crucial to manage the resource demands.

Presence of outliers. We have presented two algorithms, DAUG for data augmentation and UFDM for unification. In both algorithms, the presence of outliers may occur at two stages, one, at the raw data stage, and second, in the outputs of the involved processing techniques.

Raw ordinal data, which is bounded by a few labels, leaves no scope for outliers. However, numerical attributes are prone to outliers, though this could be handled using normalization techniques, such as Z-score. If we have ordinal values we may use the median centering instead of mean [19].

However, in the second stage, the K-means algorithm is prone to outliers. It is well known that K-means performs inappropriately when there are outliers present and is sensitive to their existence. A robust multi-view K-means method with outlier detection to remove the class outliers and attribute outliers can be applied [13]. To inherit the effectiveness of the classical K-means algorithm, with a low time complexity these methods are applied. This is the direction that this work will take going forward.

However, it is crucial to remember that the effectiveness of these strategies depends on choosing the techniques for the appropriateness of the augmentation and unification methods used.

6. Conclusion

In this work, we proposed our approach into two steps and applied the K-means clustering algorithm at each step. We first apply the augmentation technique to generate enough instances to incorporate the richness of data. We measure the deviation of augmented data from original data with descriptive statistical measures. The results are compared for every attribute so that we can employ our method suitable for considering the whole population. The overlapping region for most attributes

demonstrates that the enhanced dataset does not diverge much from the original data. This technique allows access to the huge population of the expanded dataset. The inferences we draw from the augmentation process should also apply to larger survey sizes, however, this is the future direction of the work. Next, we performed the clustering and measured its effectiveness in every aspect. We have used many efficiency metrics for clustering. We included performance metrics like Silhouette's scores, Calinski Harabasz Index, and Silhouette Analysis Plots. The resultant outputs enhanced average and high silhouette scores for each data point show that the data points are appropriately grouped and clustered. Similar to the low Calinski-Harabasz index, the high Calinski-Harabasz index gives well-defined clusters with internally similar responses and apparent distinctions between other groups. It also indicates the distance between clusters and the compactness of each cluster. After augmentation, some numerical attributes may be present in the dataset. So in our next step, we unified the dataset and converted it into the ordinal dataset. This process also helps in generalizing the results of the inferences. After each step, we apply the K-means clustering algorithm and compare the clustering efficiency metrics at different numbers of clusters. Our proposed method shows that efficiency is improved in all such cases. At last, we come to the generalized inferences part. The outcome of both datasets is the selection of the most effective attributes for deciding whether to find the entrepreneurial competency or factors governing breast cancer. Such improved results give better inferences for decision-making. In the future, we would like to use high-dimensional attribute space.

Acknowledgments. The authors thank the reviewer(s) for their insightful comments and suggestions. The authors also express their gratitude to the Editor-in-Chief, the Editor, and the Editorial Office Assistant(s) of this journal for managing this manuscript.

Data and code availability. The program code, data, and artefacts used in this work are publicly available through the *GitHub* repository[1], necessary to run and execute for interpreting, replicating, and building on the findings reported in the paper.

Declaration of competing interest. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Aggarwal C.C.: An introduction to Cluster Analysis. In: C.C. Aggarwal, C.K. Reddy (eds.), *Data clustering. Algorithms and Applications* chapter 1, pp. 1–28, Chapman and Hall/CRC, 2018. doi: 10.1201/9781315373515-1.

[2] Agrawal T., Choudhary P.: Segmentation and classification on chest radiography: a systematic survey, *The Visual Computer*, vol. 39(3), pp. 875–913, 2023.

[1] https://github.com/Bhuppigithub/Clustering_Inferences_with_Augmentation

[3] Ahmad A., Khan S.S.: Survey of State-of-the-Art Mixed Data Clustering Algorithms, *IEEE Access*, vol. 7, pp. 31883–31902, 2019. doi: 10.1109/access.2019.2903568.

[4] Back B., Sere K., Vanharanta H.: Managing complexity in large databases using self-organizing maps, *Accounting, Management and Information Technologies*, vol. 8(4), pp. 191–210, 1998. doi: 10.1016/s0959-8022(98)00009-5.

[5] Behrend T.S., Sharek D.J., Meade A.W., Wiebe E.N.: The viability of crowdsourcing for survey research, *Behavior Research Methods*, vol. 43, pp. 800–813, 2011. doi: 10.3758/s13428-011-0081-0.

[6] Belloni A., Chernozhukov V., Hansen C.: Inference on Treatment Effects After Selection Amongst High-Dimensional Controls, *The Review Economic Studies*, vol. 81(2), pp. 608–650, 2014. doi: 10.48550/arXiv.1201.0224.

[7] Bowles C., Chen L., Guerrero R., Bentley P., Gunn R., Hammers A., Dickie D.A., Hernández M.V., Wardlaw J., Rueckert D.: GAN augmentation: Augmenting training data using generative adversarial networks, *arXiv: 181010863*, 2018.

[8] Buskirk T.D., Kirchner A., Eck A., Signorino C.S.: An Introduction to Machine Learning Methods for Survey Researchers, *Survey Practice*, vol. 11(1), pp. 1–10, 2018. doi: 10.29115/sp-2018-0004.

[9] Bzdok D., Altman N., Krzywinski M.: Statistics versus machine learning, *Nature Methods*, vol. 15, pp. 233–234, 2018. doi: 10.1038/nmeth.4642.

[10] Caliński T., Harabasz J.: A Dendrite Method for Cluster Analysis, *Communications in Statistics Theory & Methods*, vol. 3(1), pp. 1–27, 1974. doi: 10.1080/03610927408827101.

[11] Cameron A.C., Miller D.L.: A Practitioner's Guide to Cluster-Robust Inference, *Journal Human Resources*, vol. 50(2), pp. 317–372, 2015. doi: 10.3368/jhr.50.2.317.

[12] Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P.: SMOTE: synthetic minority over-sampling technique, *Journal Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. doi: 10.1613/jair.953.

[13] Chen C., Wang Y., Hu W., Zheng Z.: Robust multi-view K-means clustering with outlier removal, *Knowledge-Based Systems*, vol. 210, 106518, 2020. doi: 10.1016/j.knosys.2020.106518.

[14] Chen Y., Tang S., Bouguila N., Wang C., Du J., Li H.: A fast clustering algorithm based on pruning unnecessary distance computations in DBSCAN for high-dimensional data, *Pattern Recognition*, vol. 83, pp. 375–387, 2018. doi: 10.1016/j.patcog.2018.05.030.

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

[16] Dempster A.P., Laird N.M., Rubin D.B.: Maximum likelihood from incomplete data via the EM algorithm, *Journal Royal Statistical Society: Series B (Methodological)*, vol. 39(1), pp. 1–22, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x.

[17] van Dyk D.A., Meng X.L.: The Art of Data Augmentation, *Journal of Computational and Graphical Statistics*, vol. 10(1), pp. 1–50, 2001. doi: 10.1198/10618600152418584.

[18] Firdaus S., Uddin M.A.: A survey on clustering algorithms and complexity analysis, *International Journal of Computer Science Issues*, vol. 12(2), 62, 2015.

[19] García-Jara G., Protopapas P., Estévez P.A.: Improving Astronomical Time-series Classification via Data Augmentation with Generative Adversarial Networks, *The Astrophysical Journal*, vol. 935(1), 23, 2022. doi: 10.3847/1538-4357/ac6f5a.

[20] Giordan M., Diana G.: A clustering method for categorical ordinal data, *Communications in Statistics-Theory & Methods*, vol. 40(7), pp. 1315–1334, 2011. doi: 10.1080/03610920903581010.

[21] Golinko E., Sonderman T., Zhu X.: CNFL: Categorical to Numerical Feature Learning for Clustering and Classification. In: *2017 IEEE Second International Conference on Data Science in Cyberspace (DSC), Shenzhen, China*, pp. 585–594, IEEE, 2017. doi: 10.1109/DSC.2017.87.

[22] Graubardand B.I., Korn E.L.: Inference for Superpopulation Parameters using Sample Surveys, *Statistical Science*, vol. 17(1), pp. 73–96, 2002. doi: 10.1214/ss/1023798999.

[23] He H., Bai Y., Garcia E.A., Li S.: ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong*, pp. 1322–1328, IEEE, 2008. doi: 10.1109/IJCNN.2008.4633969.

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

[25] Kim K., Hong J.S.: A hybrid decision tree algorithm for mixed numeric and categorical data in regression analysis, *Pattern Recognition Letters*, vol. 98, pp. 39–45, 2017. doi: 10.1016/j.patrec.2017.08.011.

[26] Kumar B., Kumar R.: Difference-Attribute-Based Clustering for Ordinal Survey Data. In: A.K. Dubey, V. Sugumaran, P.H.J. Chong (eds.), *Advanced IoT Sensors, Networks and Systems. SPIN 2022*, pp. 17–27, Springer, Singapore, 2022. doi: 10.1007/978-981-99-1312-1_2.

[27] Kumar B., Kumar R.: Entropy-based clustering for subspace pattern discovery in ordinal survey data. In: V. Bhateja, X.S. Yang, J. Chun-Wei Lin, R. Das (eds.), *Intelligent Data Engineering and Analytics. FICTA 2022. Smart Innovation, Systems and Technologies*, pp. 509–519, Springer, Singapore, 2022. doi: 10.1007/978-981-19-7524-0_45.

[28] Kumar B., Kumar R.: Unification of Numerical and Ordinal Survey Data for Clustering-based Inferencing, *INFOCOMP Journal Computer Science*, vol. 22(1), 2023. https://infocomp.dcc.ufla.br/index.php/infocomp/article/view/2492.

[29] Kumar R., Rockett P.: Multiobjective genetic algorithm partitioning for hierarchical learning of high-dimensional pattern spaces: a learning-follows-decomposition strategy, *IEEE Transactions on Neural Networks*, vol. 9(5), pp. 822–830, 1998. doi: 10.1109/72.712155.

[30] Ley C., Martin R.K., Pareek A., Groll A., Seil R., Tischer T.: Machine learning and conventional statistics: making sense of the differences, *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 30(3), pp. 753–757, 2022. doi: 10.1007/s00167-022-06896-6.

[31] Luchi D., Rodrigues A.L., Varejão F.M.: Sampling approaches for applying DBSCAN to large datasets, *Pattern Recognition Letters*, vol. 117, pp. 90–96, 2019. doi: 10.1016/j.patrec.2018.12.010.

[32] Mamabolo M.A., Myres K.: A detailed guide on converting qualitative data into quantitative entrepreneurial skills survey instrument, *The Electronic Journal of Business Research Methods*, vol. 17(3), pp. 102–117, 2019. doi: 10.34190/JBRM.17.3.001.

[33] Mason M.: Sample size and saturation in PhD studies using qualitative interviews, *Forum: Qualitative Sozialforschung/Forum: Qualitative Social Research*, vol. 11(3), 2010. doi: 10.17169/fqs-11.3.1428.

[34] Nardo M.: The quantification of qualitative survey data: a critical assessment, *Journal Economic Surveys*, vol. 17(5), pp. 645–668, 2003. doi: 10.1046/j.1467-6419.2003.00208.x.

[35] Pakhira M.K.: A Linear Time-Complexity k-Means Algorithm Using Cluster Shifting. In: *2014 International Conference on Computational Intelligence and Communication Networks, CICN'2014*, pp. 1047–1051, IEEE, 2014. doi: 10.1109/CICN.2014.220.

[36] Rastogi R., Mondal P., Agarwal K., Gupta R., Jain S.: GA based clustering of mixed data type of attributes (numeric, categorical, ordinal, binary, and ratio-scaled), *BIJIT – BVICAM's International Journal of Information Technology*, vol. 7(2), pp. 861–866, 2015.

[37] Rich T.S.: South Korean perceptions of unification: Evidence from an experimental survey, *Georgetown Journal of International Affairs*, vol. 20, pp. 142–149, 2019. doi: 10.1353/gia.2019.0022.

[38] Rodriguez M.Z., Comin C.H., Casanova D., Bruno O.M., Amancio D.R., Costa L.d.F., Rodrigues F.A.: Clustering algorithms: A comparative approach, *PloS one*, vol. 14(1), e0210236, 2019. doi: 10.1371/journal.pone.0210236.

[39] Rousseeuw P.J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis, *Journal of Computational & Applied Mathematics*, vol. 20, pp. 53–65, 1987. doi: 10.1016/0377-0427(87)90125-7.

[40] Sadh R., Kumar R.: Clustering of Quantitative Survey Data based on Marking Patterns, *INFOCOMP Journal Computer Science*, vol. 19(2), pp. 109–119, 2020.

[41] Sadh R., Kumar R.: Transformation and classification of ordinal survey data, *Computer Science*, vol. 24(2), 2023. doi: 10.7494/csci.2023.24.2.4871.

[42] Schliep E.M., Hoeting J.A.: Data augmentation and parameter expansion for independent or spatially correlated ordinal data, *Computational Statistics & Data Analysis*, vol. 90, pp. 1–14, 2015. doi: 10.1016/j.csda.2015.03.020.

[43] Stevens S.S.: On the theory of scales of measurement, *Science*, vol. 103(2684), pp. 677–680, 1946. doi: 10.1126/science.103.2684.677.

[44] Taylor L., Nitschke G.: Improving Deep Learning with Generic Data Augmentation. In: *2018 IEEE Symposium Series on Computational Intelligence (SSCI), Bangalore, India*, pp. 1542–1547, IEEE, 2018. doi: 10.1109/SSCI.2018.8628742.

[45] Temraz M., Keane M.T.: Solving the class imbalance problem using a counterfactual method for data augmentation, *Machine Learning with Applications*, vol. 9, 100375, 2022. doi: 10.1016/j.mlwa.2022.100375.

[46] Tourangeau R.: Cognitive aspects of survey measurement and mismeasurement, *International Journal of Public Opinion Research*, vol. 15(1), pp. 3–7, 2003. doi: 10.1093/ijpor/15.1.3.

[47] Valsiner J., Molenaar P.C., Lyra M.C.D.P., Chaudhary N.: *Dynamic Process Methodology in the Social and Developmental Sciences*, Springer, New York, 2009.

[48] Van Hulse J., Khoshgoftaar T.M., Napolitano A.: Experimental perspectives on learning from imbalanced data. In: *ICML '07: Proceedings of the 24th International Conference on Machine Learning*, pp. 935–942, Association for Computing Machinery, New York, 2007. doi: 10.1145/1273496.1273614.

[49] Velleman P.F., Wilkinson L.: Nominal, ordinal, interval, and ratio typologies are misleading, *The American Statistician*, vol. 47(1), pp. 65–72, 1993. doi: 10.1515/9783110887617.161.

[50] Zhang Y., Cheung Y.M.: Learnable Weighting of Intra-Attribute Distances for Categorical Data Clustering with Nominal and Ordinal Attributes, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44(7), pp. 3560–3576, 2021. doi: 10.1109/TPAMI.2021.3056510.

[51] Zhang Y., Cheung Y.M., Tan K.C.: A Unified Entropy-Based Distance Metric for Ordinal-and-Nominal-Attribute Data Clustering, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31(1), pp. 39–52, 2019. doi: 10.1109/TNNLS.2019.2899381.

Affiliations

Bhupendera Kumar
Jawaharlal Nehru University, Data to Knowledge (D2K) Lab, School of Computer & Systems Sciences, New Delhi 110 067, India, bkchauhan86@gmail.com

Rajeev Kumar
Jawaharlal Nehru University, Data to Knowledge (D2K) Lab, School of Computer & Systems Sciences, New Delhi 110 067, India, rajeevkumar.cse@gmail.com

Received: 14.08.2023
Revised: 29.12.2023
Accepted: 08.01.2024