Computer Science • 24(2) 2023 https://doi.org/10.7494/csci.2023.24.2.4871

Roopam Sadh
Rajeev Kumar

TRANSFORMATION AND CLASSIFICATION OF ORDINAL SURVEY DATA

Abstract *Currently, machine learning is being significantly used in almost all of the research domains; however, its applicability in survey research is still in its infancy. In this paper, we attempt to highlight the applicability of machine learning in survey research while working on two different aspects in parallel. First, we introduce a pattern-based transformation method for ordinal survey data. Our purpose for developing such a transformation method is two-fold: first, our transformation facilitates the easy interpretation of ordinal survey data and provides convenience while applying standard machine-learning approaches; and second, we demonstrate the application of various classification techniques over real and transformed ordinal survey data and interpret their results in terms of their suitability in survey research. Our experimental results suggest that machine learning coupled with a pattern-recognition paradigm has tremendous scope in survey research.*

Keywords machine learning, classification, transformation, ordinal data, survey research

Citation Computer Science 24(2) 2023: 205–224

1. Introduction

Survey research is quite popular in several domains of importance; e.g., education, health, psychology, policy science, organizational research, etc. [13]. It attempts to explain the phenomenon under consideration by analyzing the opinions of a large population [2]. The popular analysis approach in survey research relies mainly on the statistical modeling of the survey error, which requires the relational mapping of the outcomes and the covariates to be known a priori [46]. However, this criterion is not always achievable, as the functional form of such relationships is not always available for complex real-world problems [24]. Such complex scenarios require flexible modeling approaches that do not demand prior definitions of the relational mappings. These scenarios can be described better if the relational mappings can be defined based on the inherent features of the data. For example, categorizing data points based on some sort of natural proximity could lead to a better understanding of a phenomenon; e.g., the behavior analysis of a sampled population is one such application.

In addition to the requirement of flexible modeling techniques, certain data-related aspects should also be handled with care in order to gain meaningful and reliable insights from survey data. Survey data has some distinct characteristics such as heterogeneity, ordered relationships, and the significance of category labels [37]. Heterogeneity implies that data may contain several types of measurements such as binary, continuous, categorical, and their combinations [41]. Such heterogeneous data poses serious challenges in the analysis phase [32]. Furthermore, small ordinal measurements are highly prevalent in survey data – especially in the case of web-based survey applications [15]. The value-based treatment of a small ordinal scale is considered to be inappropriate in the literature [30]. In contrast, pattern-based analysis approaches perform better over such ordinal-valued vectors [47]. Respondent category labels are also of high significance in survey applications, as relationships among the variables are generally sought out concerning these categories [36].

Considering the factors that are mentioned above (the requirement of flexible modeling techniques and the distinct features of survey data), we envisage the vast potential of machine learning (ML) in the field of survey research. In contrast to model-driven statistics, ML is basically data-driven. It works without a prior understanding of the relationship between the data and the outcomes [28]. In other words, it offers flexible modeling techniques that define the relationships between the data and the outcomes purely on the basis of the inherent features of the data [10]. Second, the use of ML could lead to adding a new dimension of generalizable predictive modeling in survey research, which has been limited thus far to drawing population inferences from a sample [11]. Classification, for instance, has some direct applications in surveys, as it captures a user-defined notion of grouping data points by using a model that is trained on previously categorized data objects [1]. Some of the major applications of classification in surveys include the validation of theories, the extraction of unique patterns, behavior mapping, etc. [17]. Since the survey data is heterogeneous, it may also require transformations in many applications depending on the

objectives; e.g., categorical to continuous conversion (and vice-versa) [33, 34]. Such a transformation will also be needed for the easy interpretation and descriptions of the inherent properties of the survey data.

Therefore, we focus on two different aspects in parallel in this paper and describe them together to elaborate on their significance. First, we introduce a pattern-based data-transformation method for ordinal survey data that works based on vector arithmetic. The objective behind the development of such a transformation is to exploit the efficacy of the pattern-based approach to describe the data in more-convenient and -graphical ways. Our proposed transformation projects survey observations inside a three-dimensional cylindrical data space. It transforms an ordinal-valued survey observation that has a comparatively large set of dimensions into a three-dimensional real-valued vector. Such a visualization ability facilitates interpreting the natural features of data more conveniently.

Second, we apply several classifiers over a real survey data set and evaluate their performance in order to test their suitability for survey applications. In this way, we made primary inferences regarding the utility of the ML paradigm in survey research. One other objective behind choosing the classification in this paper is to verify the effectiveness of our proposed transformation method as to whether (and to what extent) it retains the natural features of the data. For doing so, we transform the original survey data set by our proposed pattern-based transformation method and subsequently apply the same set of classifiers. The results of the classification and the performance comparison of the different classifiers (over original and transformed data) suggest that our proposed transformation method works quite well on ordinal survey data. Thus, we can say that our transformation method is suitable for survey applications. Even though the classifiers predict the labels of unseen observations quite accurately on behalf of a model that is trained over transformed data, we observe that the transformation causes a fair amount of information loss (which impacts the performance of the classifiers). However, such information loss is a common issue in transformation methods – especially those that reduce the dimensions of the data [22]. Thus, future research is needed to handle this issue more effectively and make the transformation method more robust.

Paper organization: Section 2 describes our motivation behind developing our transformation method. Section 3 provides a brief review of the literature that describes the need for the transformation and classification of survey data. Section 4 presents the concept of the proposed pattern-based transformation method. The classification of the survey data and the corresponding transformed data is presented in Section 5. The advantages and limitations of the proposed transformation method are discussed in Section 6. Finally, Section 7 concludes the paper.

2. Motivation

Surveys generally use very few marking levels; e.g., four to seven Likert levels [26]. The objective behind this is to indicate the differences in respondents' choices and,

therefore, the magnitude of the differences has little to do here (whereas the number of differences is quite a significant factor). For this reason, the magnitude-based treatment of such ordinal survey data infers no meaningful information [36]. In contrast, analyzing marking patterns can infer more about respondents' opinions. Two respondents can be said to share similar opinions if their marking patterns are similar even if their marking values vary. For describing this phenomenon, we depict an example data set that has five observations in Figure 1 along with their trends of marking. We assume eight variables (V1, V2, ... V8) and a five-level marking scale in the example.

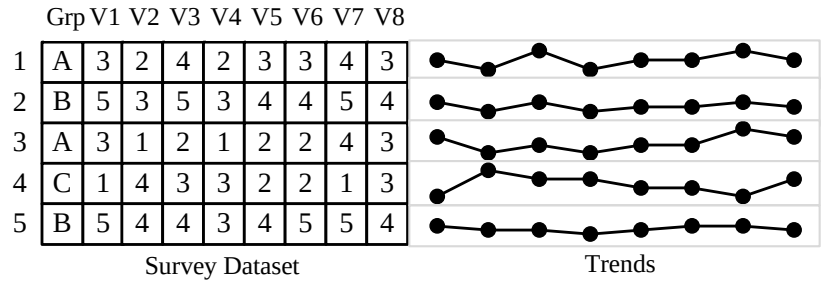

Figure 1. Example survey sample with five observations

We can observe in Figure 1 that Observations 2 and 5 look quite similar to each other, Observations 3 and 4 are proximate in some dimensions, and 1 is quite different than the others when taking the magnitudes into account. When we turn our eyes to the patterns of marking depicted nearby, however, we can find that Observations 1, 2, and 3 are quite similar; meanwhile, Observations 4 and 5 are different than the others pattern-wise. Supposing the usual survey scenario in which several observations that have several variables can exist, we can anticipate the problem of analyzing the similarities among respondents' opinions at a gross level.

The problem can be better analyzed if there is some possibility to visualize the data points in 3-D visual data space. Since no such visualization method exists to address this issue, the only way that remains is to analyze such ordinal data on the basis of magnitudes; e.g., central tendency and spread. Therefore, one of our major objectives in this paper is to design a transformation method that can visualize ordinal survey observations in a real-valued 3-D visual data space.

3. Related work

Since we focus on two different aspects in this paper, we divided this section into two subsections: type conversion in survey data analysis, and the utility of machine learning in survey research.

3.1. Type conversion in survey data analysis

In several situations, data needs to be re-scaled or transformed into other data types prior to an analysis [33]. Generally, the conversion of continuous data into a discrete form (qualitizing) requires some systematic grouping criterion, while the conversion of discrete data into a continuous form (quantitizing) needs some numerical scoring method [40]. However, such types of conversion might suffer from issues of subjectivity and information loss [27].

3.1.1. Qualitizing

Qualitizing refers to the process of transforming quantitative data into qualitative data. Such a transformation can be done through five different types of qualitative profiling schemes: modal, average, comparative, normative, and holistic [42]. Modal profiling gives descriptions to groups according to the most frequent attributes. The average profiling describes a group according to the mean of an attribute. Comparative profiling describes a group based on a comparison of its members on one or more sets of scores. Normative profiling compares the group members on behalf of one or more than one instrument; however, both the comparative and normative profiling schemes are based on the quantitative clustering of members. In contrast, holistic profiling is done based on impressions instead of scoring (or attributes), but it might include a combination of other profiling schemes [40].

3.1.2. Quantitizing

Quantitizing is the process of transforming qualitative data into quantitative data. In this process, verbal or visual data is generally reduced into an item, variable, or construct. Such a transformation is made through a narrative analysis [5]. Michela Nardo presented a mathematical approaches for quantifying macroeconomic survey data, which included probability-based methods, a time-varying parameter method, and a regression-based conversion approach [33]. The common subjective probability distribution-based conversion method was introduced by Theil [44]. This method was later extended by Knobl and Carlson & Parking, which is popularly known as the Carlson-Parking (CP) method [12, 25]. The time-varying parameters method is an extension of the CP method in which the indifference interval is allowed to vary over time while the assumption of symmetry is dropped [33]. The regression method utilizes the relationship between actual values and the respondents' perceptions as a way to quantify future expectations [35].

3.2. Utility of machine learning in survey research

Statistical modeling serves two purposes: explaining data (explanatory modeling), and predicting outcomes (predictive modeling) [9]. Modeling is supposed to approximate and incorporate various design issues (e.g., non-response, non-coverage, etc.) in a functional form. However, prior knowledge about the true function is not always available and estimating the potentially complex function in a parametric framework

might be infeasible. Machine learning provides flexible modeling that requires no prior functional representation. Moreover, machine-learning techniques have the capability to represent the complex non-linear and non-additive interrelations between outcomes and covariates [24]. Therefore, machine learning has significant potential in survey applications. The regression of continuous covariates, classification, and the clustering of categorical data are some of the potential applications where machine learning can provide effective solutions [10].

Though recent trends in computer science and other relevant fields show significant use of machine learning, its application in survey data analysis is still in its infancy. The literature suggests that survey scientists are exploring new dimensions in survey research through machine learning. For example, Christoph Kern provided an introduction of tree-based supervised learning methods and their utility in survey research [24]. The use of machine learning can also be seen in the field of developmental economics [6]. The least absolute shrinkage and selection operator (LASSO) and its adaptations (such as debiasing principle-based LASSO and hierarchical LASSO) are in use for making inferences from survey data [3, 4, 45]. For example, hierarchical LASSO was used to explore survey data and make causal inferences [7]. The low-dimensional projection estimator (LDPE) was used to explore the determinants of infant malnutrition and the effectiveness of government interventions [8, 48]. Similarly, the idea of directional pattern-based clustering has been proposed quite recently for the effective cluster analysis of ordinal survey data [37]. Guided mean centroid-based clustering, directional pattern-based semi-supervised clustering, etc. are some of the apparent examples in this regard [36, 38].

4. Transformation of ordinal survey data

Data transformation aims to make data follow some desired shapes or distributions that are necessary for the applicability of standard analysis techniques. Moreover, the transformation of data may also be desired for its easy interpretation, visualization, and comparison purposes [29]. In the context of survey applications, transformation is desired to convert the format and structure of data (which effectively fits the study's objectives). For example, mixed-data analysis sometimes demands conversions of numeric variables into categorical ones (or vice-versa) [33]. Furthermore, methods that make survey data easily interpretable or visualized are always desirable, as visualization provides an easy way to study the descriptive profiles of any phenomena under study. Thus, this section presents a pattern-based data-transformation method for ordinal survey data that converts ordinal data into real-valued data. The desirable properties of the proposed method are that it works based on the mathematical assumptions of vector arithmetic and transforms ordinal data that has an arbitrary number of dimensions into a three-dimensional data space. In this way, it reduces the dimensions of the ordinal data set and allows for its straightforward and interpretable representation in 3-D space.

4.1. Proposed transformation method

The proposed method works on the pattern-based representation of ordinal data and vector arithmetic. It works in two phases: (i) the creation of directional-difference patterns, and (ii) the vectorized representation of the patterns.

4.1.1. Creation of directional difference pattern

The proposed transformation first converts survey observations into patterns of directional differences. A directional-difference pattern is an array of the directional values (-1, 0, and $+1$) that correspond to a survey observation that represents the relative significance of each variable with respect to its preceding neighbor. For example, if a variable is larger than the preceding variable, its directional value is represented by $+1$; if it is smaller, then its corresponding directional value is represented by -1 (while 0 represents equal variable values). The value of the first variable is defined by comparing it with half of the maximum marking scale (n/2 for odd scale and (n+1)/2 for even scale, where n denotes the used marking levels). Figure 2 depicts the first example observation of our example survey sample that is described in Section 2 (Figure 1) (along with its corresponding directional-difference pattern).

Grp	V1	V2	V3	V4	V5	V6	V7	V8
A	3	2	4	2	3	3	4	3

A	1	-1	1	-1	1	0	1	-1

Figure 2. First sample survey observation with its directional-difference pattern

4.1.2. Vectorized representation of patterns

After converting the survey observations into patterns of directional differences, the proposed transformation method observes each such pattern in a *vectorized data space*. Before going into the details of our proposed vectorized data space, let us recall the concept of a resultant vector. Suppose a point object at the coordinate position (x, y) of the Cartesian system on which f_1 and f_2 forces of m_1 and m_2 magnitudes act in different directions. The directions of these forces are defined with respect to the x-axis such that force f_1 is θ_1 degrees apart and force f_2 is θ_2 degrees apart from the horizontal x-axis. In this scenario, the resultant force (vector) R of magnitude M_R and θ_R degrees apart from the x-axis can be calculated by the vector summation of both of these forces. Figure 3 represents the assumed scenario.

Mathematically, we can express forces f_1, f_2, and resultant force R (depicted by Figure 3) in their vector form as follows:

$$f_1 = (m_1 \cos\theta_1)i + (m_1 \sin\theta_1)j$$

$$f_2 = (m_2 \cos\theta_2)i + (m_2 \sin\theta_2)j$$

Then, $R = f_1 + f_2$, and

$$M_R = \sqrt{(X_R - x)^2 + (Y_R - y)^2}$$
$$\theta_R = 180 - \arctan\left(\frac{Y_R}{X_R}\right)$$

where :
$X_R = (m_1 \cos\theta_1 + m_2 \cos\theta_2)$
$Y_R = (m_1 \sin\theta_1 + m_2 \sin\theta_2)$

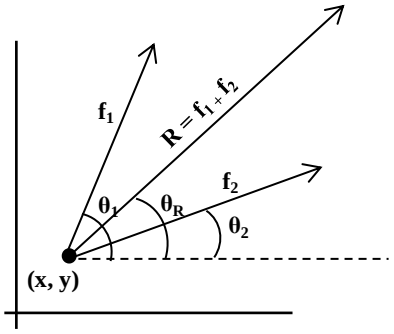

Figure 3. Demonstration of resultant vector

Our transformation method observes each pattern of directional difference in a three-dimensional space (x, y, z). Since the pattern is defined on behalf of three equidistant direction values $(-1, 0, +1)$, the method divides the z dimension into two halves (positive and negative) centered around the origin. A positive z-axis measures the number of positive direction values (+1s) in the pattern, and a negative z-axis denotes the number of negative direction values (−1s) in the pattern. The x–y planes on both the negative and positive z halves are divided into the number of parts defined by the total number of variables in the pattern. The x–y plane in the positive half (z-positive) corresponds only to the positive values (+1s) in the pattern and treats the other direction values (0 and −1) as zero. The same is true for the negative half that corresponds only to the negative direction values (−1s) and treats the other values as zero. The correspondence between the negative and positive x–y planes is that they are 180 degrees apart angularly from each other. Figure 4 describes the positive and negative x–y planes for a pattern that has eight direction values (the eight survey variables that are denoted as V1 to V8 in the observation).

A pattern is seen as two separate points in the negative and positive halves depending on the numbers of +1s and −1s. Let us take our example pattern of a length of eight (Figure 2) that has four +1s, three −1s, and one 0. Then, the positive x–y plane will appear at 4 on the z-axis, and the negative x–y plane will appear at −3 on the z-axis. The position of the points in the negative and positive x–y planes will be calculated through vector arithmetic. Consider each direction value as a unit force that acts upon a point object at the origin and take the case of the positive x–y plane in the example.

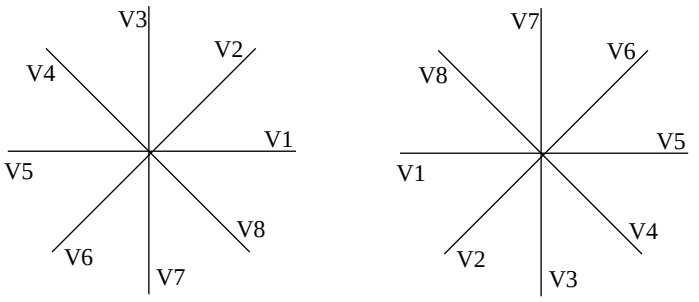

Figure 4. x–y planes on positive and negative z-axes

So, the first, third, fifth, and seventh variables can be considered to be active unit forces that act on the point at the origin for the positive half. The rest of the variables are considered to be dormant (asserting no force on the point). The position of the point in the x–y plane will be seen as the resultant force that is calculated by the vector addition (described in the previous paragraphs) of the four unit-forces (V1, V3, V5, V7) that act in the directions that are shown in Figure 4. Since Force V5 works in the opposite direction of Force V1, it cancels out the effects of V1. The same is true in the cases of V3 and V7, where they nullify the effects of each other. This means the point remains at the origin of the positive x–y plane, as all forces combinedly cancel out the effects of each other. Now consider the negative x–y plane of Figure 4 and our example pattern where Unit Forces (Variables) V2, V4, and V8 act on the point object at its origin. These variables are 225° (V2), 315° (V4), and 135° (V8) apart from the x-axis. The resultant force that is generated by the combined effect of these forces is of a unit magnitude, and it shifts the point object at position $(-0.7, -0.7)$ in the negative x–y plane. The resultant is 225° apart from the x-axis, which is the direction of Force V2. By combining the z-axis in these coordinates of the positive and negative x–y planes, we can locate the points in 3-D space. Thus, our example pattern can be shown in 3-D space with two points that are described by coordinates $X_1 = 0$, $Y_1 = 0$, $Z_1 = 4$ and $X_2 = -0.7$, $Y_2 = -0.7$, $Z_2 = -3$.

In this way, a directional pattern in our defined data space can be described by a line segment that connects the two separate points – one in the positive half that is defined by the positive x–y plane, and one in the negative half that is defined by the negative x–y plane. The length of the line segment is dependent on the z-axis, which defines the number of 1s in the positive half and the number of -1s in the negative half. Thus, a pattern in the above design can be represented by a line segment inside a cylindrical data space. For instance, our example pattern is shown as the dotted line in the middle of the cylindrical data space that is shown in Figure 5.

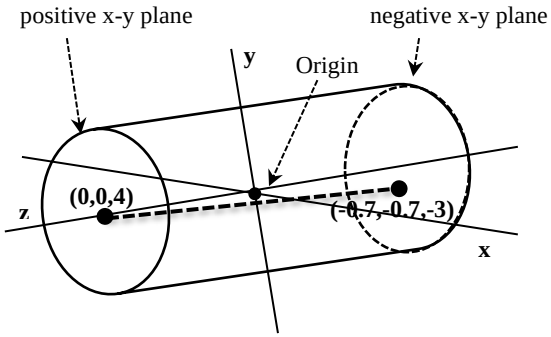

Figure 5. Example pattern inside our proposed 3-D cylindrically shaped data space

In this manner, our transformation method converts each survey observation into an array of six-coordinate values. First, it converts the survey observation into the pattern of the directional differences; then, it observes each pattern in a string of six real values (denoting the coordinates of two points in our defined 3-D vectorized space). Such a transformation is advantageous for two reasons. First, it makes the visualization of data points easy, as each survey observation can be visualized as a line segment inside the cylindrical data space that is depicted in Figure 5. Thus, the properties of the data points can be described based on line parameters such as slope and intercept. Second, the method converts the ordinal valued observations into real-valued vectors; thus, making the analysis methods that are designed for real values applicable on the ordinal data sets that are transformed through our method.

Let us take our example survey sample that is described in Section 2. We described that Observations 2 and 5 looked similar and also that Observations 3 and 4 looked slightly similar to each other when taking the magnitudes into account. According to the survey application point of view (pattern-based similarity), however, Observations 1, 2, and 3 show similar respondents' preferences despite their diverse magnitudes. We now apply our transformation method over our example survey sample and observe their proximity in 3-D space. Figure 6 depicts the actual ordinal survey sample and its corresponding real-valued transformed observations.

In Figure 6, we can observe that Observations 1, 2, and 3 are placed at the same location in the transformed data space (showing their pattern-wise similarity), whereas Observations 4 and 5 are placed at different locations (showing their pattern-wise dissimilarity). In this way, our transformation method made it easy to interpret the sample data by visualizing it in 3-D space. One added advantage with our transformation is that it converts data that has arbitrary dimensions (8-D in the example case) to six-dimensional real valued data (the end points of a line segment inside a 3-D cylindrical data space). This visualization capability of our transformation method is quite helpful in the perspective of a descriptive analysis of ordinal survey data.

	Grp	V1	V2	V3	V4	V5	V6	V7	V8
1	A	3	2	4	2	3	3	4	3
2	B	5	3	5	3	4	4	5	4
3	A	3	1	2	1	2	2	4	3
4	C	1	4	3	3	2	2	1	3
5	B	5	4	4	3	4	5	5	4

Example Dataset

	Grp	X1	Y1	Z1	X2	Y2	Z2
1	A	0	0	4	-.7	-.7	-3
2	B	0	0	4	-.7	-.7	-3
3	A	0	0	4	-.7	-.7	-3
4	C	1.4	0	2	0	0	-4
5	B	-.7	-.7	3	-.7	.7	-3

Transformed Dataset

Figure 6. Example survey sample along with transformed replica

5. Classification of ordinal survey data

Classification is a significant problem in the field of knowledge discovery. The aim of classification is to learn the structure of an available data set by fitting the model and use this model to predict the category of unseen data instances [1]. In the context of survey applications, classification can be used to predict trends, describe respondents' marking behaviors, validate theories, etc. In this section, we apply different classifiers over a survey data set that was collected during a study that was aimed at exploring the significance of the quality parameters of higher educational institutions (HEIs) [39]. First, we give a brief description of the data set along with its basic statistics; then, we proceed to make primary inferences based on the results of the chosen classifiers that are applied over the data set. We then transform the data by using our proposed transformation method and apply the same set of classifiers over transformed data. We compare the classifiers' performances over the original and transformed data to describe the utility and robustness of the proposed transformation method in the context of survey applications. Specifically, we want to test how many inherent characteristics of the data that our transformation method can retain despite the information loss due to the dimensionality reduction.

5.1. Used data set

The data set was collected during our previous study, which was intended to explore the various quality parameters of HEIs [39]. Eleven quality parameters were explored in this study through the grounded theory method. The data (qualitative as well as quantitative) was collected from the National Capital Region (NCR) of India. The NCR was chosen since it contained representative premier institutions, and the population in these institutions represented all of India. Seven respondent categories were identified in the survey: undergraduate (UG), graduate studies (GS), graduate research (GR), faculty (FAC), parents (PAR), administrators (AD), and professionals (PRO). Since the proportion of the administrator category was imbalanced, we excluded the administrators' responses. In total, 2533 responses were considered for an exploratory analysis; of these, 438 responses were from the UG category, 463 from

GS, 447 from GR, 401 from FAC, 395 from PAR, and 389 from PRO. The basic statistics of these eleven quality parameters (mean values and standard deviations) are provided in Table 1.

Table 1
Survey items and their corresponding mean values and standard deviations

Sr	Survey items	Mean	SD
1.	Teaching	3.2	0.74
2.	Graduate Outcomes	3.0	0.89
3.	Academic Flexibility	3.0	0.76
4.	Transparency & Accountability	3.0	0.77
5.	Infrastructure & Resources	3.0	0.73
6.	Research	3.0	0.87
7.	Student Support Services	2.9	0.78
8.	International Outlook	2.8	0.81
9.	Fee & Financial Assistance	2.7	0.96
10.	Academic Autonomy	2.5	0.92
11.	Inclusivity	2.5	0.84

In this study, we will treat the 11 survey items that were contained in the data as independent variables and the respondent category label as a dependent variable (classes) for classification. We will evaluate the results in light of the stakeholder theory, as it is well-established in the higher education domain; we also utilized it while exploring the quality parameters (survey) [19, 31]. A detailed description of the stakeholder theory and its applicability in higher education can be found in the work of Jongbloed [23]. The higher education domain is associated with several internal and external stakeholders that have significantly divergent perceptions regarding educational quality [18]. Since the perceptions of academic stakeholders are divergent with respect to academic quality, this phenomenon should also be reflected by educational survey data; therefore, we hypothesize that the classification of educational survey data should identify each stakeholder class according to the inherent features of the survey data set.

5.2. Classification of original survey data

We used the *Waikato Environment for Knowledge Analysis* – popularly known as the WEKA platform (Version 3.8.5) – for classifying our data set. WEKA provides a rich set of machine-learning techniques for data analysis and predictive modeling [20]. We chose ten classifiers – two each from five categories (the tree-based, Bayes theorem-based, function-based, ensemble-based, and instance-based classifiers) for classifying our data set. Our chosen classifiers included the random forest, random tree, random subspace, random committee, logistic, multilayer perceptron, naive Bayes, BayesNet, IBK, and KStar classifiers. We used a ten-fold cross-validation test procedure and standard (default) parameter settings for each of the chosen classifiers.

Since our data set contained ordinal-valued responses regarding 11 survey items that were structured in a simple data grid, we first passed our original data set (categorical form) as input to the classification methods. A summary of the results for the chosen classifiers is presented in Table 2. This summary presents the weighted averages of the chosen assessment parameters for the classified respondent categories by the classifiers. We chose seven popular classification-assessment parameters: false-positive rate (FPR), accuracy (ACC), precision (PRE), recall (REC), F-measure (F-M), the area under the receiver operating characteristic curve (ROC), and the area under the precision recall curve (PRC) [43].

Since each performance measure has its own strengths and limitations depending on the context and the type of data to which it is applied [21], we present the results of each of the chosen eight in the tables just to present a holistic profile of the classifier's performance. For clarity and readability purposes, however, we chose ROC in our text discussions most of the time for comparison.

Table 2
Performance of various classifiers over original data (ordinal)

Classifier	FPR	ACC.	PRE.	REC.	F-M.	ROC	PRC
Random Forest	0.003	0.995	0.986	0.986	0.986	0.999	0.997
Random Tree	0.005	0.992	0.978	0.977	0.977	0.990	0.971
Random Subspace	0.019	0.969	0.908	0.906	0.906	0.991	0.963
Random Committee	0.003	0.995	0.985	0.985	0.985	0.999	0.999
Logistic	0.062	0.899	0.693	0.696	0.693	0.914	0.749
Multilayer Perceptron	0.018	0.970	0.911	0.910	0.909	0.960	0.923
Naive Bayes	0.067	0.888	0.661	0.665	0.660	0.895	0.698
BayesNet	0.068	0.888	0.659	0.664	0.659	0.895	0.698
IBK	0.002	0.996	0.989	0.989	0.989	0.999	0.996
KStar	0.002	0.996	0.989	0.989	0.989	1.000	0.999

A summary of the results that are presented in Table 2 suggests that the chosen classifiers identified the existing classes in our data quite accurately. Most of the classifiers surpassed a score of 0.90 out of the maximum 1.0 for ROC and the other measures (ACC, PRE, REC, F-M, and PRC). The ROC values for random forest, random tree, random committee, and IBK were all 0.99 (almost 1.0), which suggests that these classifiers predicted accurate labels for nearly all of the test samples. These trends signify that each category in the data had its own specific relational pattern in the data set. Even though the results of a few of the classifiers (Logistic, naive Bayes, and BayesNet) were slightly inferior, they were still quite impressive in general (ROC – greater than 0.89). The raw implication of these results is that ML classification techniques are quite good at extracting patterns from ordinal survey data like ours, and tree-based, ensemble-based, and instance-based classification methods can work exceptionally well in such scenarios. However, generalizing this implication is a topic

for further research, and the classification methods need to be evaluated on several other ordinal survey data sets to confirm such an implication.

These results can also be interpreted from the point of view of the application of ML in survey methodology. On the basis of these classification results, we can validate our primary hypothesis (validating the stakeholder theory). These classification results signify that the existing classes in data have their specific inherent structures. This phenomenon suggests that each academic stakeholder category has quite peculiar opinions regarding the quality of HEIs. The existing respondent categories (academic stakeholders) are easily recognizable based on their choice patterns; based on the specificity of their opinions, they can be predicted with utmost confidence. Our classification results thus validated the applicability of the stakeholder theory in the domain of the quality of HEIs. Even though this phenomenon is well-known in the academic domain from a theoretical point of view, these classification results serve as solid empirical evidence of the existence of such a phenomenon. These results thus affirm that machine learning is quite useful in analyzing complex social constructs and can serve several important applications of survey research.

On several occasions, ordinal data (especially in self-administered online surveys) has been treated as real-valued [30]. Thus, we attempt to analyze the effects of the magnitude-based treatment of ordinal values over classifier performances. We passed our original data set as a real-valued data set to classifiers and observed their results (ordered categories were mapped into consecutive integers 1, 2, 3, and 4). A summary of the classification results over such converted data is given in Table 3.

Table 3
Performance of various classifiers over original data (real)

Classifier	FPR	ACC.	PRE.	REC.	F-M.	ROC	PRC
Random Forest	0.003	0.995	0.984	0.984	0.984	0.999	0.997
Random Tree	0.004	0.994	0.982	0.982	0.982	0.991	0.973
Random Subspace	0.025	0.959	0.881	0.878	0.877	0.985	0.945
Random Committee	0.003	0.995	0.985	0.985	0.985	0.999	0.997
Logistic	0.078	0.871	0.614	0.615	0.614	0.872	0.653
Multilayer Perceptron	0.058	0.907	0.729	0.720	0.722	0.869	0.661
Naive Bayes	0.072	0.879	0.632	0.637	0.627	0.883	0.680
BayesNet	0.071	0.882	0.643	0.647	0.642	0.888	0.691
IBK	0.003	0.996	0.988	0.988	0.988	0.996	0.990
KStar	0.003	0.995	0.987	0.987	0.987	1.000	0.999

The gross results (for the ROC score) of all of the classifiers except for the tree-based, instance-based, and ensemble-based classifiers showed a slight decline in their performance after ordinal-to-real conversion. The observable decline in ROC could be found to correspond with the function-based and Bayes theorem-based classifiers (logistic, multilayer perceptron, BayesNet, and naive Bayes). Even though the differences in the results (Table 2 and Table 3) were not very significant, this tells us that

the magnitude-based treatment of ordinal values impacted the relational structure of the survey data. In this way, these results validate the previous studies that have emphasized avoiding the magnitude-based treatment of ordinal data [16, 30]. However, future research is needed to solidify this aspect empirically.

5.3. Classification of transformed survey data

In this subsection, we attempt to test the robustness of our transformation technique by applying the chosen classifiers over transformed data and analyzing the predictive power of the models. Since our transformation method reduces ordinal data of any dimension into six-dimensional real-valued vectors, it causes a significant amount of information loss. The information loss in our transformation occurs at two levels: one – at the time of converting original data into directional patterns, and second – at the time of converting direction vectors into the end-points of the line segments. Such a conversion makes descriptive analyses and interpretations of data easier, as the data-points can be easily visualized inside a cylindrical coordinate system. However, reducing the dimensions while preserving the natural features of the data is a trade-off [22]. A transformation method (especially one that reduces dimensions) is said to be effective if it can balance these two contradictory aspects. Therefore, we hypothesize that our transformation method will be useful if it can retain a fair amount of predictive power despite the information loss. Our assumption is that, if classifiers that are applied over data (transformed by our method) can predict the true class labels of test samples with sufficient accuracy, then it can be said that our transformation method is useful and robust in the context of survey applications.

We first transform original survey data by applying our proposed transformation method. The transformation of original survey data converts 11-dimensional ordinal observations into 6-dimensional real-valued vectors. These six-dimensional vectors depict the coordinates of the three-dimensional end-points of the line segments inside the proposed vectorized data space. In this way, our transformation method significantly reduces the size of the data (an $\approx .45\%$ reduction in size). Such a size reduction is beneficial in many aspects. The reduced data is easier to handle, as it requires fewer processing resources. In addition, a descriptive analysis of reduced information is easy, as it can be visualized in a 3-D data space.

We applied classifiers over transformed data and evaluated their performance based on the chosen assessment parameters. A summary of the classifiers' performance over the transformed data is given in Table 4. It can be observed from Table 4 that, except for the function-based and Bayes theorem-based classifiers (logistic, multilayer perceptron, naive Bayes, and BayesNet), all of the other classifiers performed remarkably well over the transformed data. Their ROC (as well as ACC) values, which were greater than 0.97, suggest that they can easily identify classes and accurately predict class labels for unseen test observations based on their patterns. Although the performance of the classifiers declined slightly over the transformed data, the results were pretty impressive and optimistic. These results suggest that our transformation

method could retain a sufficiently good amount of data features despite the information loss. Thus, we can say that our proposed transformation method works well for ordinal survey data and is quite useful for survey applications. The declines in the classifiers' performances were attributed to the significant reduction of the data size. Such a reduction causes a loss of information regarding the interrelationships among the variables to some extent. However, this is a common problem with data-reduction methods, and it is an open-ended research problem [14]. Moreover, the proposed transformation has some limitations: it only applies to ordinal data that has small scales, and it cannot handle high-dimensional data.

Table 4
Performance of various classifiers over transformed data

Classifier	FPR	ACC.	PRE.	REC.	F-M.	ROC	PRC
Random Forest	0.017	0.973	0.920	0.919	0.919	0.994	0.978
Random Tree	0.017	0.971	0.915	0.915	0.915	0.979	0.935
Random Subspace	0.032	0.948	0.847	0.844	0.843	0.974	0.911
Random Committee	0.016	0.974	0.922	0.921	0.921	0.993	0.978
Logistic	0.135	0.775	0.301	0.325	0.305	0.666	0.303
Multilayer Perceptron	0.116	0.809	0.418	0.426	0.417	0.721	0.403
Naive Bayes	0.127	0.790	0.359	0.371	0.356	0.675	0.329
BayesNet	0.092	0.850	0.553	0.549	0.550	0.858	0.608
IBK	0.016	0.973	0.919	0.919	0.919	0.977	0.934
KStar	0.017	0.973	0.918	0.918	0.918	0.994	0.979

6. Discussion

Survey research has an important place in several key domains. The conventional survey paradigm relies mainly on the statistical modeling of survey errors, which requires that the relationships between the variables and the outcomes be defined in a functional form. However, this is not always achievable in various real-world problems. Machine learning is the most suitable approach for such scenarios. Thus, this paper revolves around the utility of machine learning in survey research by taking the example of classification and its utility in analyzing complex social constructs. We applied various classification methods to a real educational survey data set and evaluated their results with respect to the well-established stakeholder theory. The results suggested that the classification of the used survey data satisfied the stakeholder theory quite well, which depicts the potential of machine learning in the survey research domain. In addition, we also introduced a data-transformation method that converts ordinal survey data into real-valued vectors. Our transformation method enables ordinal observations to be represented as line segments inside a 3-D cylindrical Cartesian system, thus making its descriptive analysis relatively easier. To show the strength of the proposed concept, we transformed the used survey data through

the proposed method and evaluated the performance of various classifiers over the transformed data. The classification results suggested that the proposed transformation method worked quite well on the used survey data. However, we also observed that the transformation might cause significant information loss – especially if the data is high-dimensional. Moreover, the proposed method was limited to small-scale ordinal values; thus, future research should be directed toward making a more robust and general-purpose data-transformation method in order to eliminate such problems.

7. Conclusion

This paper outlined the requirements and potential of machine learning in survey research. The article dealt with two crucial aspects – the transformation and classification of ordinal survey data. We introduced a transformation method that converts ordinal survey data into a six-dimensional continuous feature vector. Our transformation method observes each survey observation inside a 3-D cylindrical data space, making its interpretation easier. We also applied various classifiers to real survey data and its corresponding transformed replica. We evaluated the performance of the chosen classifiers on both original and transformed data. The results suggested that most of the classifiers that were trained over our data predicted the actual class labels with high accuracy. Such optimistic results affirm that the proposed transformation method works quite well over ordinal survey data and that it is quite suitable for survey applications.

References

[1] Aggarwal C.C.: Data classification. In: *Data Mining. The Textbook*, pp. 285–344, Springer, Cham, 2015.

[2] Behrend T.S., Sharek D.J., Meade A.W., Wiebe E.N.: The viability of crowdsourcing for survey research, *Behavior Research Methods*, vol. 43(3), pp. 800–813, 2011.

[3] Belloni A., Chernozhukov V., Hansen C.: Inference on treatment effects after selection among high-dimensional controls, *The Review of Economic Studies*, vol. 81(2), pp. 608–650, 2014.

[4] Bien J., Taylor J., Tibshirani R.: A lasso for hierarchical interactions, *Annals of Statistics*, vol. 41(3), 1111, 2013.

[5] Borkan J.M., Quirk M., Sullivan M.: Finding meaning after the fall: injury narratives from elderly hip fracture patients, *Social Science & Medicine*, vol. 33(8), pp. 947–957, 1991.

[6] Brahma D.: *Essays in the Application of Machine Learning in Development Economics*, Western Michigan University, 2019.

[7] Brahma D., Mukherjee D.: India's Universal Immunization Program: a lesson from Machine Learning, *Economics Bulletin*, vol. 39(1), pp. 581–591, 2019.

[8] Brahma D., Mukherjee D.: Infant malnutrition, clean-water access and government interventions in India: a machine learning approach towards causal inference, *Applied Economics Letters*, vol. 28(16), pp. 1426–1431, 2021.

[9] Breiman L.: Statistical modeling: The two cultures (with comments and a rejoinder by the author), *Statistical Science*, vol. 16(3), pp. 199–231, 2001.

[10] 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.

[11] Bzdok D., Altman N., Krzywinski M.: Statistics versus machine learning, *Nature Methods*, vol. 15(4), pp. 233–234, 2018.

[12] Carlson J.A., Parkin M.: Inflation expectations, *Economica*, vol. 42(166), pp. 123–138, 1975.

[13] Church A.H., Waclawski J., Kraut A.I.: *Designing and Using Organizational Surveys: A Seven-Step Process*, Business and Management Series, Jossey-Bass, San Francisco, 2001.

[14] DeMers D., Cottrell G.W.: Non-linear dimensionality reduction. In: *Advances in neural information processing systems*, pp. 580–587, Citeseer, 1993.

[15] Göb R., McCollin C., Ramalhoto M.F.: Ordinal methodology in the analysis of Likert scales, *Quality & Quantity*, vol. 41(5), pp. 601–626, 2007.

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

[17] Grice J.W.: Observation oriented modeling: preparing students for research in the 21st century, *Comprehensive Psychology*, vol. 3, pp. 1–27, 2014.

[18] Harvey L., Green D.: Defining quality, *Assessment & Evaluation in Higher Education*, vol. 18(1), pp. 9–34, 1993.

[19] Harvey L., Williams J.: Fifteen Years of Quality in Higher Education, 2010.

[20] Holmes G., Donkin A., Witten I.H.: Weka: A machine learning workbench. In: *Proceedings of ANZIIS'94-Australian New Zealnd Intelligent Information Systems Conference*, pp. 357–361, IEEE, 1994.

[21] Japkowicz N., Shah M.: *Evaluating learning algorithms: a classification perspective*, Cambridge University Press, 2011.

[22] Johansson S., Johansson J.: Interactive dimensionality reduction through user-defined combinations of quality metrics, *IEEE transactions on visualization and computer graphics*, vol. 15(6), pp. 993–1000, 2009.

[23] Jongbloed B., Enders J., Salerno C.: Higher education and its communities: Interconnections, interdependencies and a research agenda, *The Future of Higher Education and the Future of Higher Education Research*, vol. 56(3), pp. 303–324, 2008.

[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] Knöbl A.: Price expectations and actual price behavior in Germany, *Staff Papers*, vol. 21(1), pp. 83–100, 1974.

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

[27] Leon de A.R., Chough K.C.: *Analysis of Mixed Data: Methods & Applications*, CRC Press, 2013.

[28] Ley C., Martin R.K., Pareek A., Groll A., Seil R., Tischer T.: Machine learning and conventional statistics: making sense of the differences, 2022.

[29] Manikandan S.: Data transformation, *Journal of Pharmacology and Pharmacotherapeutics*, vol. 1(2), 126, 2010.

[30] Merbitz C., Morris J., Grip J.C.: Ordinal scales and foundations of misinference, *Archives of Physical Medicine and Rehabilitation*, vol. 70(4), pp. 308–312, 1989.

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

[32] Morgan J.N., Sonquist J.A.: Problems in the analysis of survey data, and a proposal, *Journal American Statistical Association*, vol. 58(302), pp. 415–434, 1963.

[33] Nardo M.: The quantification of qualitative survey data: a critical assessment, *Journal of Economic Surveys*, vol. 17(5), pp. 645–668, 2003.

[34] Onwuegbuzie A.J., Combs J.P.: Data analysis in mixed research: A primer, *International Journal of Education*, vol. 3(1), 2011.

[35] Pesaran M.H., Weale M.: Survey Expectations, *Handbook of Economic Forecasting*, vol. 1, pp. 715–776, 2006.

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

[37] Sadh R., Kumar R.: Clustering of quantitative survey data: A subsystem of EDM framework. In: *Computational Methods and Data Engineering*, pp. 307–319, Springer, 2021.

[38] Sadh R., Kumar R.: Directional Pattern based Clustering for Quantitative Survey Data: Method and Application, *Survey Research Methods*, vol. 15(2), pp. 169–185, 2021.

[39] Sadh R., Kumar R.: Dimensional Inadequacy of Rankings: Exploring Substantial and Meta-quality Dimensions for Higher Educational Institutions, *Academia*, vol. 26, pp. 25–48, 2022.

[40] Sandelowski M.: Combining qualitative and quantitative sampling, data collection, and analysis techniques in mixed-method studies, *Research in Nursing & Health*, vol. 23(3), pp. 246–255, 2000.

[41] Stevens S.S.: On the Theory of Scales of Measurement, *Science*, vol. 103(2684), pp. 677–680, 1946.

[42] Tashakkori A., Teddlie C., Teddlie C.B.: *Mixed Methodology: Combining Qualitative and Quantitative Approaches*, vol. 46, sage, 1998.

[43] Tharwat A.: Classification assessment methods, *Applied Computing and Informatics*, 2020.

[44] Theil H.: On the time shape of economic microvariables and the Munich business test, *Revue de l'Institut International de Statistique*, pp. 105–120, 1952.
[45] Tibshirani R.: Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58(1), pp. 267–288, 1996.
[46] Tourangeau R.: Cognitive aspects of survey measurement and mismeasurement, *International Journal of Public Opinion Research*, vol. 15(1), pp. 3–7, 2003.
[47] Valsiner J., Molenaar P.C., Lyra M.C., Chaudhary N.: *Dynamic Process Methodology in the Social and Developmental Sciences*, Springer, 2009.
[48] Zhang C.H., Zhang S.S.: Confidence intervals for low dimensional parameters in high dimensional linear models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76(1), pp. 217–242, 2014.

Affiliations

Roopam Sadh
Jawaharlal Nehru University, School of Computer & Systems Sciences, New Delhi – 110067, India, roopam.sadh@gmail.com

Rajeev Kumar
Jawaharlal Nehru University, School of Computer & Systems Sciences, New Delhi – 110067, India, rajeevkumar.cse@gmail.com

Received: 03.06.2022
Revised: 17.12.2022
Accepted: 20.12.2022