Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 2, 2022 (New Series), pp 15-25

Symbolic Data Analysis *vs* Classical Data Analysis: A Comparative Study

Dipanka Bora[1] and Hemanta Saikia[2]
[1]Teaching Associate of Agricultural Statistics, FGI College of Agricultural Sciences, Hengbung, Manipur
[2]Hemanta Saikia, Assistant Professor of Statistics, College of Sericulture, Assam Agricultural University, Jorhat-13, Assam

Received: 28 April 2021; Revised: 07 June 2021; Accepted: 09 June 2021

Abstract

A symbolic data set is a combination of symbolic values. The analysis of these symbolic values is known as symbolic data analysis. It is an extension of the standard classical data analysis where symbolic data tables are used as input and symbolic objects are made output as a result. Symbolic data may arise in all branches of science and social science after aggregating a base data set over individual entries that together constitute a category of interest. This study attempts to bring into notice the use of symbolic data analysis and compare its outcome with standard classical data analysis. Different statistical tools have been used for comparative analysis of the symbolic and classical data *viz.* descriptive statistics, covariance, and correlation. To apply these statistical tools in both symbolic and classical data analysis set up, a well-known *Iris* flower data set is being used. The outcome of the study shows that there is a little difference in the results of descriptive statistics for the univariate case between classical data analysis and symbolic data analysis. However, in bivariate statistics computation though the directions of the covariance and correlation values (*i.e.* positive or negative) are the same, yet symbolic data analysis gives comparatively lesser magnitude values than the classical data analysis.

Key words: Data analysis; Descriptive statistics; Interval-valued variables; Symbolic data.

1. Introduction

When we deal with classical data set that data may be either univariate or bivariate or multivariate. In the case of univariate classical data, a single random variable is considered (*e.g.* production of rice). For bivariate data, two random variables (*e.g.* amount of fertilizer and production of wheat) are studied simultaneously in respect of their distribution. Similarly, more than two random variables concerning their distributions for a multivariate classical data are considered (*e.g.* monthly information of temperature, rainfall, humidity, *etc.*). Usually, classical data analysis seeks to describe the descriptive statistics and determine the reliability of inferential statistics. It is based on the repeatedly measured properties of the same objects or only one value per object.

Statistically, classical data on $\boldsymbol{P}$ random variables are represented by a single point (say) in $\boldsymbol{P}$-dimensional space. For instance, the observed values for the random variable $Y = (Y_1, Y_2, \ldots, Y_p)$ for a single individual. This type of data can be analyzed using classical techniques for $n = 150$ observations (say) with $\boldsymbol{P} = 30$ variables. When the size of n becomes

Corresponding Author: Dipanka Bora
Email: dipankabora666@gmail.com

very large (*e.g.* $\boldsymbol{n}$ = 50000 and $\boldsymbol{P}$ = 90), standard classical analysis can be knotty. Again, consider a random variable "Type of Disease ($\boldsymbol{Y}$)" with $\boldsymbol{Y}$ = {Diabetes, High Blood Pressure, Cancer} then a classical response from the respondent could be $\boldsymbol{Y}$ = Diabetes, or $\boldsymbol{Y}$ = High Blood Pressure, or $\boldsymbol{Y}$ = Cancer. It can be noticed that each observation consists of only one value or data point. Now if a respondent has two diseases say Diabetes and High Blood Pressure [*i.e.*$\boldsymbol{Y}$= {Diabetes, High Blood Pressure}] then the typical classical data set format can't accommodate this information. In such situations, symbolic data analysis can be ready to lend a hand. However, there are two possible issues. The first one is how the data set can be prepared to a size that allows analysis to proceed appropriately. The second issue is, to attain the first one, it is essential to consider what we want to learn or extract from the given data set. Symbolic data may arise in all branches of science and social science (*e.g.* from medical, industry, government experiments, and other data collection pursuits) in a variety of different ways. It may arise after aggregating a base data set over individual entries that together constitute a category of interest to the researcher (Diday and Fraiture, 2008). Furthermore, they may arise as an outcome of aggregating very large data set into a smaller manageable sized data set or aggregating into a data set that provides information about categories of interest (Diday and Fraiture, 2008). More specifically, we could say that a symbolic value typically represents the set of individuals who satisfy the description of the associated symbolic concept or category. A symbolic value may include lists, intervals, categories and so on. A more elaborate discussion of symbolic values is provided in the following section.

Symbolic data analysis is an extension of standard data analysis where symbolic data tables are used as input and symbolic objects are made output as a result. The data units are called symbolic since they are more complex than standard ones, as they donot only contain values or categories but also include internal variation and structure (Billard and Diday, 2006). Suppose we have a data set that can be structured like a classical data set. This data can be aggregated to a manageable size and categorized with specializing decision to construct symbolic data sets. Like classical data, symbolic data set can also have three types of variables *viz.* interval-valued, multi-valued and modal variables. The multi-valued variables are the different attributes of the symbolic data set which can have a relation with other variables. Suppose for a field experiment on rice, presence and absence of fungal disease, and the number of spraying for treatment can be jointly considered as multi-valued variables. The interval-valued variable has the maximum and minimum value of the observation, where the values of the observation are varied. The modal variables are multistate variables with a frequency, probability or weight attached to a specific value in the data. Usually, these weights are capacities, creditabilities, necessities or possibilities.

Now let us explain how a symbolic dataset is created from the classical data set. The Table 1 consists of a set of classical data which represents the different varieties of rice along with the information of season, production, tillers per hill, duration and grain size.

As mentioned, a symbolic value may be lists, intervals, categories, *etc.*; from Table 1, the season variable may be considered as a concept to construct a symbolic data set. It could be described by considering rice season (*i.e.* Sali, Ahu, and Bodo) as the concept. The set of seasons is the extent and the different characteristics of rice are the intent. Thus, using the different seasons of rice as a concept a symbolic data table is constructed (*cf.* Table 2). In Table 2, the variable production, number of tillers, and durations are interval-valued variables. For these interval-valued variables, other variables vary within the respective symbolic values. Here except grain size, all the variables are quantitative interval-valued variables. The variable grain size is qualitative *viz.* Big or Small. To transform this qualitative

variable into symbolic values, first, we calculated the ratio of big and small grains corresponding to the rice season. Thereafter, these ratio values are assigned as symbolic values to the variable grain size to make the variable as an interval-valued variable.

Table 1: Classical data set of rice varieties

Variety	Season	Production (kg/ha)	Tillers/hill (number)	Duration (days)	Grain Size
Ranjit	Sali	75	25	155	Big
Kushal	Sali	65	20	150	Big
Satyaranjan	Sali	54	18	135	Small
Lachit	Ahu	42	13	115	Small
Luit	Ahu	36	11	105	Small
Silarai	Ahu	45	15	125	Big
Bishnuprasad	Bodo	66	21	160	Big
Jyotiprasad	Bodo	60	19	170	Big
Joymoti	Bodo	78	26	175	Small

(*Source: Leaflet of Regional Agricultural Research Station, Titabar, Assam Agricultural University, Jorhat*)

The following symbolic data table obtained using the season as a concept from Table 1.

Table 2: Symbolic data set of rice varieties

Season	Production (kg/ha)	Tillers/hill (number)	Duration (days)	Grain Size
Sali	[54,75]	[18,25]	[135,155]	(0.67B, 0.33S)
Ahu	[36,45]	[11,15]	[105,125]	(0.33B, 0.67S)
Bodo	[60,78]	[19,26]	[160,175]	(0.67B, 0.33S)

From the above discussion, we understand that classical values can be qualitative or quantitative. In contrast, symbolic values can be single-valued, interval-valued, and multi-valued with or without logical dependency rules. However, we have especially focused on interval-valued variables in this study. These days the researchers are more acquainted with the classical data and its modeling, so the importance of symbolic data analysis is always quarantined. Therefore, this study attempts to bring into notice the symbolic data analysis and compare its outcome with classical data analysis. Different statistical tools have been used for comparative analysis of symbolic as well as classical data *viz.* descriptive statistics, covariance, and correlation.

2. Review of Literature

Statistical data analysis always plays an important role in determining useful and effective information on real-life situations. In the words of Tukey (1962), data analysis is the "*Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyze data*." In an era of big data, the prominence of symbolic data analysis is

indispensable through which one could summarize big data into smaller data set of manageable size. Several authors have already added some value to the literature through their contribution.

Symbolic objects are the basic elements for knowledge representation in symbolic data analysis (Prediger, 1997). A method of clustering for a set of symbolic data where individuals are described by symbolic variables of various types: interval, categorical multi-valued or modal variables are presented by Brito (2003). Billard and Diday (2003) summarized large datasets into a more manageable size and tried to get maximum knowledge inherent in the entire dataset as much as possible. A similar study was performed by Diday and Esposito (2003) introducing symbolic objects and constitutes an explanatory output for data analysis. Mballo and Diday (2005) studied the reliability of the Kolmogorov-Smirnov criterion to build the decision tree on interval-valued variables to extract symbolic objects from the decision tree and to induced the data table of symbolic objects for higher study of symbolic data analysis. Brito *et al*. (2006) introduced partitioning clustering methods for objects described by interval data. Appice *et al.* (2006) generalized symbolic data analysis aimed at some standard statistical data mining methods, which has developed for classification tasks in the case of symbolic objects. Brito (2007*a*) discussed some issues that arise when trying to apply classical data analysis techniques to symbolic data and addressed the vital question of the measurement of dispersion and also the result of different possible choices in the design of multivariate methods. Diday (2008) observed that databases are now ubiquitous in industrial companies and public administrations and they often grow to an enormous size. In symbolic data analysis, these categorical and numerical are considered to be the new statistical units. The next step is to get these higher-level units and then to describe them by taking care of their internal variation. Domingues *et al.* (2010) introduced a new linear regression method for interval-valued data. Fraiture *et al.* (2011) worked on symbolic data analysis and explained how the classical data models to take into account more complete and complex information. Primental *et al.* (2012) used common tools of symbolic data analysis to reduce the data without losing much information. They used information about researchers of institutions from Brazil through the tools of symbolic data. The main goal was to analyze the scientific production of Brazilian institutions. Brito (2007*b*) worked on modeling and analyzing interval data and discussed some issues that arose when applying classical data analysis techniques to interval data. She put a special focus on the notions of dispersion, association and linear combinations of interval variables and presented some methods that have been proposed for analyzing clustering, discriminant analysis, linear regression and interval time series analysis. Some Indian statisticians also worked on the field of symbolic data analysis. Dinesh *et al.* (2005) studied symbolic data analysis literature revealed that symbolic distance measures are playing a major role in solving pattern recognition and analysis problems. Guru *et al.* (2011) proposed a new model to grade cured tobacco leaves using symbolic data. Doreswamy and Narasegouda (2014) proposed an object-oriented data model using symbolic data analysis which provides a sensor data repository for storing and managing sensor data.

3. Data and Methodology

3.1. Data

The relevant data for comparative analysis between classical and symbolic data analysis is collected from the source *https://en.m.wikipedia.org/wiki/Iris_flower_data_set*. This data set is popularly known as *Iris* flower data set or Fisher's *Iris* data set. It is a multivariate data set introduced by British statistician and biologist Fisher (1936) in his paper entitled "*The*

Use of Multiple Measurements in Taxonomic Problems as an Example of Linear Discriminant Analysis". The data set consists of 50 samples from each of the three species of *Iris viz. Setosa*, *Virginica*, and *Versicolor*. The four different features were measured from each sample and they are the length and width of the sepals and petals of *Iris* flower in centimeters.

3.2. Methodology

Keeping the objective of the study in mind, the formulae of descriptive statistics, covariance and correlation are presented in the subsequent sections for symbolic data. As we believe that the formulae of the above statistical measures for classical data analysis set up are well known to the readers. Also, noticed that we have considered the methodology of symbolic data analysis for interval-valued variables only. The single-valued and multi-valued variables of symbolic data are not considered in this study for comparison.

3.2.1. Descriptive statistics for interval–valued variables

Let us define the interval-valued variable $Y_j = Z$ and the 'Z' contains the interval of 'u' number of observation. For 'u' number of observation of 'Z' interval, the values of $Z(u) = [a_u, b_u]$ for $u \in E = \{1,\ldots.., m\}$. Here a_u is the minimum value, b_u is the maximum value of the specified observation and 'm' is the total number of observations. Now for a interval-valued variable 'Z' the symbolic mean and symbolic variances are calculated by

$$\text{Symbolic Mean}\left(S_m\right) = \frac{1}{2m}\sum_{u\in E}\left(b_u + a_u\right) \tag{1}$$

$$\text{Symbolic Variance}\left(S_v\right) = \frac{1}{3m}\sum_{u\in E}\left(b_u^2 + b_u a_u + a_u^2\right) - \frac{1}{4m^2}\left[\sum_{u\in E}\left(b_u + a_u\right)\right]^2 \tag{2}$$

$$\text{Symbolic Standard Deviation } (SD_j) = \sqrt{S_v} \tag{3}$$

3.2.2. Bivariate statistics for interval–valued variables

Let $Z_1(u)$ and $Z_2(u)$ are two symbolic observations on the space $Z(u) = Z_1(u) \times Z_2(u)$. The $Z_1(u)$ contains interval symbolic variables $[a_{1u}, b_{1u}]$ at 'u' observation and $Z_2(u)$ contains interval symbolic variables $[a_{2u}, b_{2u}]$ at 'u' observation for each $u \in E$. Here 'a_{1u}' is the minimum value and 'b_{1u}' is the maximum value of Z_{1u} interval symbolic variables. Similarly, 'a_{2u}' is the minimum value and 'b_{2u}' is the maximum value of Z_{2u} interval symbolic variables. Now the symbolic covariance function between $Z_1(u)$ and $Z_2(u)$ interval-valued symbolic variables is defined as

$$Cov\left(Z_1, Z_2\right) = \frac{1}{4m}\left\{\sum_{u\in E}\left(b_{1u} + a_{1u}\right)\left(b_{2u} + a_{2u}\right)\right\} - \frac{1}{4m^2}\left\{\sum_{u\in E}\left(b_{1u} + a_{1u}\right)\right\}\left\{\sum_{u\in E}\left(b_{2u} + a_{2u}\right)\right\} \tag{4}$$

Once we have a covariance function, we can easily calculate the symbolic correlation between the interval–valued variables Z_1 and Z_2. It is defined as

$$r\left(Z_1, Z_2\right) = \frac{Cov\left(Z_1, Z_2\right)}{\sqrt{Var\left(S_{v1}\right)} \times \sqrt{Var\left(S_{v2}\right)}} \tag{5}$$

where $r(Z_1, Z_2)$ is the symbolic correlation between interval-valued variables Z_1 and Z_2, variance (S_{v1}) is the symbolic variance of $Z_1(u)$ and variance (S_{v2}) is the symbolic variance of $Z_2(u)$.

4. Results and Discussion

A classical data set is a group of contents of a single database table where every column of the table represents a particular variable and each row corresponds to a given member of the data set. On the other hand, the symbolic data set is a combination of symbolic values *viz.* intervals, lists, categories and so on. We have constructed the symbolic data set from the classical data set in case of the univariate and bivariate case using the concept of interval-valued symbolic values. To do that statistical language *R* is being used with package *RSDA* (*cf.* Appendix-A) and the corresponding symbolic data is presented in Table 3.

Table 3: Symbolic data set of *Iris* flower data

Species	Sepal Length $[Z_1(u)]$	Sepal Width $[Z_2(u)]$	Petal Length $[Z_3(u)]$	Petal Width $[Z_4(u)]$
	$[a_{1u}, b_{1u}]$	$[a_{2u}, b_{2u}]$	$[a_{3u}, b_{3u}]$	$[a_{4u}, b_{4u}]$
Setosa	[4.3, 5.8]	[2.3, 4.4]	[1.0, 1.9]	[0.1, 0.6]
Versicolor	[4.9, 7.0]	[2.0, 3.4]	[3.0, 5.1]	[1.0, 1.8]
Virginica	[4.9, 7.9]	[2.2, 3.8]	[4.5, 6.9]	[1.4, 2.5]

Table 3 represented the symbolic data set of the classical *Iris* flower data set, which is having interval-valued variables. The variables *viz.* Sepal length, Sepal width, Petal Length and Petal width in the table contain [minimum value, maximum value] corresponding to the number of observations or species. In this symbolic data set, the factor species is considered as a concept and accordingly we have four symbolic variables are $Z_1(u)$, $Z_2(u)$, $Z_3(u)$ and $Z_4(u)$.

4.1. Univariate statistics of classical and symbolic data for *Iris* flower data set

The descriptive statistics for classical data analysis of *Iris* flower data set are presented to compare with the symbolic data analysis. We have used the usual statistical tools to calculate the descriptive statistics presented in Table 4.

Table 4: Descriptive statistics of *Iris* flower data set for classical data analysis

Descriptive Statistics	Sepal Length (Y_1)	Sepal Width (Y_2)	Petal Length (Y_3)	Petal Width (Y_4)
Mean	5.8433	3.0573	3.7580	1.1933
Variance	0.6856	0.1899	3.1162	0.5810
Standard deviation	0.8280	0.4358	1.7653	0.7622

Likewise, to calculate the descriptive statistics for symbolic data analysis of *Iris* flower data set, some of the basic and essential computations we have to perform at first. Thereafter, the symbolic mean (S_m), symbolic variance (S_v) and symbolic standard deviation (S_{sd}) are calculated using the formula given in equation (1), (2) and (3) respectively. In Table 5, the total values for each of the four variables are obtained by $\sum_{i=1}^{4}(b_{iu} + a_{iu})$ and

$\sum_{i=1}^{4}\left[b_{iu}^2+\left(b_{iu}\times a_{iu}\right)+a_{iu}^2\right]$. Now if we look at the values of descriptive statistics from both the classical and symbolic data analysis table, the values are not differing too much. It means that symbolic data analysis gives almost the same mean, variance and standard deviation values which we have computed from the classical data analysis for *Iris* flower data.

Table 5: Descriptive statistics of *Iris* flower data set for symbolic data analysis

Sepal Length [$Z_1(u)$]	a_{1u}	b_{1u}	$b_{1u}+a_{1u}$	$[b_{1u}^2+(a_{1u}\times b_{1u})+a_{1u}^2]$	S_m	S_v	S_{sd}
Setosa	4.3	5.8	10.1	77.07	5.8	0.75	0.86
Versicolor	4.9	7.0	11.9	107.31			
Virginica	4.9	7.9	12.8	125.13			
Total			34.8	309.51			
Sepal Width [$Z_2(u)$]	a_{2u}	b_{2u}	$b_{2u}+a_{2u}$	$[b_{2u}^2+(a_{2u}\times b_{2u})+a_{2u}^2]$	S_m	S_v	S_{sd}
Setosa	2.3	4.4	6.7	34.77	3.02	0.32	0.56
Versicolor	2.0	3.4	5.4	22.36			
Virginica	2.2	3.8	6.0	27.64			
Total			18.1	84.77			
Petal Length [$Z_3(u)$]	a_{3u}	b_{3u}	$b_{3u}+a_{3u}$	$[b_{3u}^2+(a_{3u}\times b_{3u})+a_{3u}^2]$	S_m	S_v	S_{sd}
Setosa	1.0	1.9	2.9	6.51	3.73	3.37	1.83
Versicolor	3.0	5.1	8.1	50.31			
Virginica	4.5	6.9	11.4	98.91			
Total			22.4	155.73			
Petal Width [$Z_4(u)$]	a_{4u}	b_{4u}	$b_{4u}+a_{4u}$	$[b_{4u}^2+(a_{4u}\times b_{4u})+a_{4u}^2]$	S_m	S_v	S_{sd}
Setosa	0.1	0.6	0.7	0.43	1.23	0.49	0.76
Versicolor	1.0	1.8	2.8	6.04			
Virginica	1.4	2.5	3.9	11.71			
Total			7.4	18.18			

4.2. Bivariate statistics of classical and symbolic for *Iris* flower data set

For comparing the bivariate statistics between classical and symbolic data set of *Iris* flower data, we computed the covariance and correlation between the variables. The results are presented in the following tables.

Table 6: Bivariate statistics of *Iris* flower data set for classical data analysis

Bivariate Statistics	Y_1Y_2	Y_1Y_3	Y_1Y_4	Y_2Y_3	Y_2Y_4	Y_3Y_4
Covariance	-0.0422	1.2658	0.5123	-0.3275	-0.1205	1.2854
Correlation	-0.1176	0.8718	0.8179	-0.4284	-0.3654	0.9627

The symbolic covariance between Sepal Length $Z_1(u)$ and Sepal Width $Z_2(u)$, Sepal Length $Z_1(u)$ and Petal Length $Z_3(u)$, *etc.* for interval-valued symbolic variables are obtained

by using the equation (4). Likewise, we have calculated all the possible symbolic covariance and correlation between the variables and the results can be seen in Table 7.

In Table 7, column-wise the $Z_1(u) \times Z_2(u)$ represents the symbolic covariance (-0.1025) and symbolic correlation (-0.2097) of Sepal Length and Sepal Width, the $Z_1(u) \times Z_3(u)$ represents the symbolic covariance (0.98) and symbolic correlation (0.6168) of Sepal Length and Petal Length, *etc.* Now let us compare the results of classical data analysis (*cf.* Table 6) and symbolic data analysis (*cf.* Table 7) from the computation of bivariate statistics. It is observed that though the directions of the values (*i.e.* positive or negative) are the same in covariance and correlation, the symbolic data analysis gives comparatively lesser magnitude values than classical data analysis. This is a result of the loss of information from the data in every step of processing. In symbolic data analysis for interval-valued variables, we deal with the maximum and minimum values of the data set instead of considering all the values like in classical real-valued data. Thus, the covariance and correlation results from symbolic data analysis have revealed lesser magnitude values in comparison to the classical data analysis results. If this is factual then the higher level of statistical analysis like multiple regression, clustering, factor analysis, *etc.* based on symbolic data might mislead the researchers to draw an appropriate inference from the data.

Table 7: Bivariate statistics of *Iris* flower data set for symbolic data analysis

Bivariate Statistics	$Z_1(u) \times Z_2(u)$	$Z_1(u) \times Z_3(u)$	$Z_1(u) \times Z_4(u)$	$Z_2(u) \times Z_3(u)$	$Z_2(u) \times Z_4(u)$	$Z_3(u) \times Z_4(u)$
Covariance	–0.1025	0.9800	0.3725	–0.2981	–0.1197	1.1597
Correlation	–0.2097	0.6168	0.6090	–0.2878	–0.3003	0.8950

For both univariate and bivariate statistics, we have computed mean, variance, and standard deviation, covariance and correlation. Though the simple linear regression analysis is not performed, the information available from univariate and bivariate statistics can easily be attained considering Y as the dependent variable and X as an independent variable in terms of classical set up like

$$\left(Y - \overline{Y}\right) = r \frac{S_y}{S_x}\left(X - \overline{X}\right) \tag{6}$$

Based on the equation (6), the equivalent symbolic linear regression equation between Sepal Length ($Z_1(u)$) and Petal Length ($Z_3(u)$) can easily be fitted. Let us consider that Sepal Length is dependent on Petal Length and then it is defined as

$$[Z_1(u) - \overline{Z}_1(u)] = r\left(Z_1, Z_3\right)\frac{S_{v1}}{S_{v3}}[Z_3(u) - \overline{Z}_3(u)] \tag{7}$$

where $r(Z_1, Z_3)$ represents the correlation and S_{v1} and S_{v3} represents the standard deviation of Sepal Length ($Z_1(u)$) and Petal Length ($Z_3(u)$) respectively.

5. Conclusion

The extension of classical exploratory data analysis to the analysis of interval-valued symbolic data raises a few pertinent questions. How to compute dispersion precisely for different types of symbolic data (*i.e.* single-valued and multi-vaued)? How to define linear combinations between the symbolic variables? Whether the properties which we have usually

considered for classical real-valued data in terms of linear models will be valid for symbolic data? It is because the way to assess the central tendency and dispersion of symbolic data is not comparable with classical real-valued data. There may be some alternatives to attain these questions which we need to explore for greater attention of the researchers across the globe. However, the choice of an alternative way shall depends on the type of symbolic data to be used subsequently. Regarding interval-valued symbolic data, the one important issue is application of statistical models. Without statistical modeling, parameter estimation and testing of hypothesis are not possible. So the challenge is in front of the researchers who wants to explore symbolic data analysis beyond the classical framework of real-valued data. In today's era of big data, where data storage and analytics is a big challenge, the exploration of symbolic data analysis in solving the problem of big data may open a new window in front of the researchers. Furthermore, the development of appropriate user-friendly statistical software to analyze the symbolic data will go a long way in tackling the challenges posed by big data.

References

Appice, A., Amato, C. D., Esposito, F. and Malebra, D. (2006). Classification of symbolic objects: A lazy learning approach. *Intelligent Data Analysis*, **10**(**4**), 301-324.

Arroyo, J. and Mate, C. (2009). Forecasting histogram time series with K-nearest neighbours methods. *International Journal of Forecasting*, **25**(**1**), 192-207.

Billard, L. and Diday, E. (2003). Symbolic Data Analysis: Definitions and examples. *Technical Report*, 62 pages, at (http://www.stat.uga.edu/faculty/LYNNE/Lynne.html).

Billard, L. and Diday, E. (2003). From the statistics of data to the statistics of knowledge: Symbolic data Analysis. *Journal of the American Statistical Association*, **98**(**462**), 470-487.

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis, Conceptual Statistics and Data Mining.* West Sussex, England: John Wiley & Sons Ltd.

Brito, P., De Carvalho, F. A. T. and Bock, H. H. (2006). Dynamic clustering for interval data based on L2 distance. *Computational Statistics*, **21**(**2**), 231-250.

Brito, P. (2003). Hierarchical and pyramidal clustering for symbolic data. *Journal of the Japanese Society of Computational Statistics*, **15**(**2**), 231-244.

Brito, P. (2007*a*). *On the Analysis of Symbolic Data*. Brito, P., Cucumel, G., Bertrand, P. and Carvalho, F. de. (Eds.) *Selected Contributions in Data Analysis and Classification* (pp. 13-22). Springer Nature, Switzerland.

Brito, P. (2007*b*). *Modelling and Analysing Interval Data*. Springer, Decker, R. and Lenz, H. J. (Eds.) (2006). *Advances in Data* Analysis. Berlin, Heidelberg.

Cury, A., Diday, E. and Cremona, C. (2010). Application of symbolic data analysis for structural modification assessment. *Engineering Structures*, **32**(**3**), 762-775.

Diday, E. and Esposito, F. (2003). An introduction to symbolic data analysis and the SODAS software. *Intelligent Data Analysis*, **7**(**6**), 583-601.

Diday, E. and Fraiture, M. (2008). *Symbolic Data Analysis and the SODAS Software*. John Wiley and Sons Ltd.: West Sussex, England.

Diday, E. and Vrac M. (2005). Mixture decomposition of distributions by copulas in the symbolic data analysis framework. *Discrete Applied Mathematics*, **147**(**1**), 27-41.

Dinesh, M. S., Gowda, K. C. and Nagabhusan, P. (2005). *Fuzzy-Symbolic Analysis for Classification of Symblic Data*. S. K. Pal (Ed.) *Pattern Recognition and Machine Intelligence: First International Conference* (pp. 338-343). Kolkata, India: Springer.

Diday, E. (2008). *Symbolic Data Analysis and the SODAS Software.* Belgium: John Wiley & Sons, Ltd.

Domingues, M. A. O, De Souza, R. M. C. R. and Cysneiors, F. J. A. (2010). A robust method for linear regression of symbolic interval data. *Pattern Recognition Letters*, **31**(**13**), 1991-1996.

Doreswamy and Narasegouda S. (2014). Symbolic data analysis for development. In *Proceeding of the International Conference on Frontiers of the Intelligent Computing: Theory and Applications* (FICTA-2013), Switzerland: Springer, 435-442.

Esposito, F., Malebra, D. and Lisi, F. (1998). Flexible matching of boolean symbolic objects. *Proc. NTTS*, **98**.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis. *Annals of Eugenics*, **7**(**2**), 179-188.

Fraiture, M. N. and Brito, P. (2011). Far beyond the classical data models: Symbolic data analysis. *Statistical Analysis and Data Mining*, **4**(**2**), 157-170.

Guru, D. S., Mallikarjuna, P. B. and Shenoi, M. M. (2011). Min-max representation of features for grading cured tobacco leaves. *Statistics and Applications*, **9**(**1&2**),15-29.

Kaytoue, M., Kuznetsov, S. O., Napoli, A. and Polaillon, G. (2011, September). Symbolic data analysis and formal concept analysis. In XVIIIeme rencontres de la Société Francophone de Classification-SFC 2011.

Lauro, C. N. and Palumbo, F. (2000). Principal component analysis of interval data: a symbolic data analysis approach. *Computational Statistics*, **15**(**1**), 73-87.

Mballo, C. and Diday, E. (2005). Decision trees on interval-valued variables. *The Electronic Journal of Symbolic Data Analysis*, **3(1)**, 1723-5081.

Nagabhusan, P. and Kumar, R. P. (2007). *Histogram PCA*. Liu Derong, H. Z. (Ed.) *Advances in Neural Networks - ISNN 2007* (pp. 1012-1022). Nanjing, China, Springer.

Prediger, S. (1997). *Symbolic Objects in Formal Concept.* Darmstadt: Frachbereich Mathematik, Technische Hochschulle.

Primental, B. A., Nobrega, J. P. and De Souza R. M. C. R. (2012). *Using Weighted Clustering and Symbolic Data to Evaluate Institutes's Scientific Production*. In Proceedings of International Conference on Artificial Neutral Networks (pp. 435-442). Springer, Berlin, Heidelberg.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, **33**(**1**), 1-67.

Appendix-A

Classical Data to Symbolic Data in *R* using the Package '*RSDA*'

```
library(RSDA)

S=classic.to.sym(data=iris, concept="Species", variables = c(Sepal.Length, Sepal.Width,
Petal.Length, Petal.Width))

S       # to get the symbolic output
```

Appendix-B

Some Basic Computation of Symbolic Data Analysis for Bivariate Statistics

Table B1: Procedure to calculate the symbolic covariance ($Z_i(u)$)

Species	$(b_{1U} + a_{1U}) \times (b_{2U} + a_{2U})$	$(b_{1U} + a_{1U}) \times (b_{3U} + a_{3U})$	$(b_{1U} + a_{1U}) \times (b_{4U} + a_{4U})$	$(b_{2U} + a_{2U}) \times (b_{3U} + a_{3U})$	$(b_{2U} + a_{2U}) \times (b_{4U} + a_{4U})$	$(b_{3U} + a_{3U}) \times (b_{4U} + a_{4U})$
Setosa	67.67	29.29	7.07	19.43	4.69	2.03
Versicolor	64.26	96.39	33.32	43.74	15.12	22.68
Verginica	76.80	145.92	49.92	68.40	23.40	44.46
Total	208.73	271.6	90.31	131.57	43.21	69.17

The symbolic covariance between Sepal Length ($Z_1(u)$) and Sepal Width ($Z_2(u)$) is calculated using the equation (4).

$$Cov(Y_1, Y_2) = \frac{1}{4\times 3}\{208.73\} - \frac{1}{4\times 3^2}[(34.8)\times(18.1)] = -0.1025 \quad \text{(B1)}$$

where, $\sum(b_{1u} + a_{2u})\times(b_{3u} + a_{4u}) = 208.73$, $\sum(b_{1u} + a_{1u}) = 34.8$ & $\sum(b_{2u} + a_{2u}) = 18.1$

Now the symbolic correlation (*cf.* equation (5)) between Sepal Length ($Z_1(u)$) and Sepal Width ($Z_2(u)$) is calculated by

$$r(Y_1, Y_2) = \frac{Cov(Y_1, Y_2)}{\sqrt{Variance(S_{v1})\times Variance(S_{v2})}} \quad \text{(B2)}$$

$$r(Y_1, Y_2) = \frac{-0.0125}{\sqrt{0.7500\times 0.3186}} = -0.2097$$

Similarly, we have calculated all the possible covariance and correlations among the variables for *Iris* flower data set using symbolic data analysis and the results are presented in Table 7.