International Journal of Chemical Studies

P-ISSN: 2349–8528
E-ISSN: 2321–4902
www.chemijournal.com
IJCS 2020; SP-8(4): 338-343
© 2020 IJCS
Received: 15-05-2020
Accepted: 19-06-2020

Sudha Bishnoi
Department of Mathematics and Statistics, CCS Haryana Agricultural University, Hisar, Haryana, India

BK Hooda
Department of Mathematics and Statistics, CCS Haryana Agricultural University, Hisar, Haryana, India

Corresponding Author:
Sudha Bishnoi
Department of Mathematics and Statistics, CCS Haryana Agricultural University, Hisar, Haryana, India

A survey of distance measures for mixed variables

Sudha Bishnoi and BK Hooda

DOI: https://doi.org/10.22271/chemi.2020.v8.i4f.10087

Abstract
Distance measures are base for many statistical and data science methods with their applicability in various fields of science. Mixed variables data which is combination of continuous and categorical variables occurs frequently in fields such as medical, agriculture, remote sensing, biology, marketing, ecology etc., but a little work has been done for evaluating distance for such type of data. As there is not much literature available on distance measures for mixed data, therefore the fundamental sources that provide a comprehensive detail of a particular measure for mixed variables data were studied and reviewed in this paper.

Keywords: distance measure, similarity measure, mixed data, heterogeneous data, k nearest neighbor, classification, discrimination

Introduction

Distance is defined as a quantitative degree of how far apart two objects are. A synonym for distance is dissimilarity. The calculation of distance between individuals or two or more groups also called populations arises in many areas such as biology, psychology, ecology, medical diagnosis and agriculture. Some statistical techniques also use the distance measures as their base like discriminant analysis, classification, clustering etc. Further distance measures are of vital importance in machine learning, they are base of many popular machine learning algorithms like k-nearest neighbor which is a supervised learning technique and k-means clustering which is an unsupervised learning technique.

When all the variables are continuous, the most commonly used distance measure is the Euclidean distance, and the simple matching coefficient is most common when all the variables are categorical. Most of the researches which need calculation of distance are confined to continuous variables, but in real world the data is mostly a combination of continuous and categorical variables also called as mixed variables data or heterogeneous data. Vast literature on distance measures is available when the data is of only continuous nature (Cha, 2007) [3] or of only categorical nature (Boriah *et al.,* 2008) [2], but when data is mix of both continuous and categorical type then most of the researchers either ignore its categorical nature and proceed with distance measures for continuous data or they transform the continuous data into categorical and proceed with distance measure for categorical data. But conversion of variables into the same scale involves loss of information.

If one wishes to retain the variables in their original form, then a reasonable solution is to develop formulae specifically for mixed data types. Gordan (1981) [7] suggested to analyze separately for each variable type and then combining those results. The various distance measures that are available for mixed type of data are explained in detail in section 2, and section 3 concludes this paper.

Distance measures for mixed variables

We begin with some basic introduction to a distance measure. Distance basically indicates how different two vectors are, it is a function which takes two input vectors and returns a real positive number called the distance between two vectors. The value of this distance function should be small between similar pointsand large between dissimilar data points. The mathematical definition of a distance measure includes three requirements to be satisfied, which are defined as:

1. The distance between two points *i* and *j* is always a value greater than or equal to zero, that is *i* is not equal to *j*, then $d_{ij} > 0$
2. The distance between *i* and *j* is equal to zero if and only if *i* is equal to *j*, that is *i* = *j*, then $d_{ij} = 0$.
3. The distance between *i* and *j* is equal to the distance between *j* and *i*, that is$d_{ij} = d_{ji}$, which implies that direction of distance measurement does not matter.

Distances which satisfy these three requirements are known as distance measures, and those distance measures which also satisfy one additional requirement are called as distance metrics. This fourth requirement is defined as:

1. Considering the presence of a third point *r*, the distance between *i* and *j* is always less than or equal to the sum of the distance between *i* and *r* and the distance between *j* and *r*, that is$d_{ij} \leq d_{ir} + d_{jr}$. This means the distance between two points cannot be larger than the sum of their distances from a third point.

A concept which is closely related to the distance measure is a similarity measure which measures the similarity of two points. It is inversely related to distance function and given as$s_{ij} = 1 - d_{ij}$.

The next most appropriate starting point of this discussion of distance measure is the Euclidean distance. Euclidean distance is the generalization of the Pythagoras theorem to many dimensions and is one of the simplest distance measures. It is the standard reference for evaluating any other distance measure, and it is the base for many distance measure derived for mixed data. Its minimum value is 0 and there is no upper bound. The Euclidean distance is calculated as:

$$d_{ij} = \sqrt{\sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^2}$$

where p is the number of variables,x_{ik} is the value *ith* observation on *kth* variable and $x_{jk}$is the value *jth*observation on *kth* variable

There are many weaknesses of this distance measure which makes it a poor choice. But many researchers still continue to useit because of its simplicity despite having many weaknesses. The Euclidean distance is popular when data is purely continuous, and next we have another popular distance measure when the data is purely categorical.

The simple matching coefficient given by Sokal & Michener, 1958 is the most widely known similarity measure for categorical variables. The distance can easily be calculated as the number of disagreements divided by the total number of variables. For observations *i* and *j* it is calculated as:

$$d_{ij} = \frac{\sum_{k=1}^{p}(x_{ik}-x_{jk})}{p}$$

where *p* is the number of variables, $x_{ik} - x_{jk} \in \{0,1\}$, $x_{ik} - x_{jk} = 0$, if $x_{ik} = x_{jk}$, and $x_{ik} - x_{jk} = 1$, if$x_{ik} \neq x_{jk}$.

Euclidean distance and matching coefficient are the most basic and popular measures for continuous and categorical data, respectively. These two measures contribute to the evolution of various distance measures for mixed data. The distance measures available for mixed variables can be categorized into two groups. First group includes the measures which are ensemble of various distance measures for different types of variables, and hence providing single distance measure for mixed variables. We named this type of distance measures as ensembled distance measures. Second group includes the distance measures which are exclusively defined for mixed type of variables, and this group is named as heterogeneous distance measures. In the figure 1, the distance measures for mixed variables are arranged in historical order.

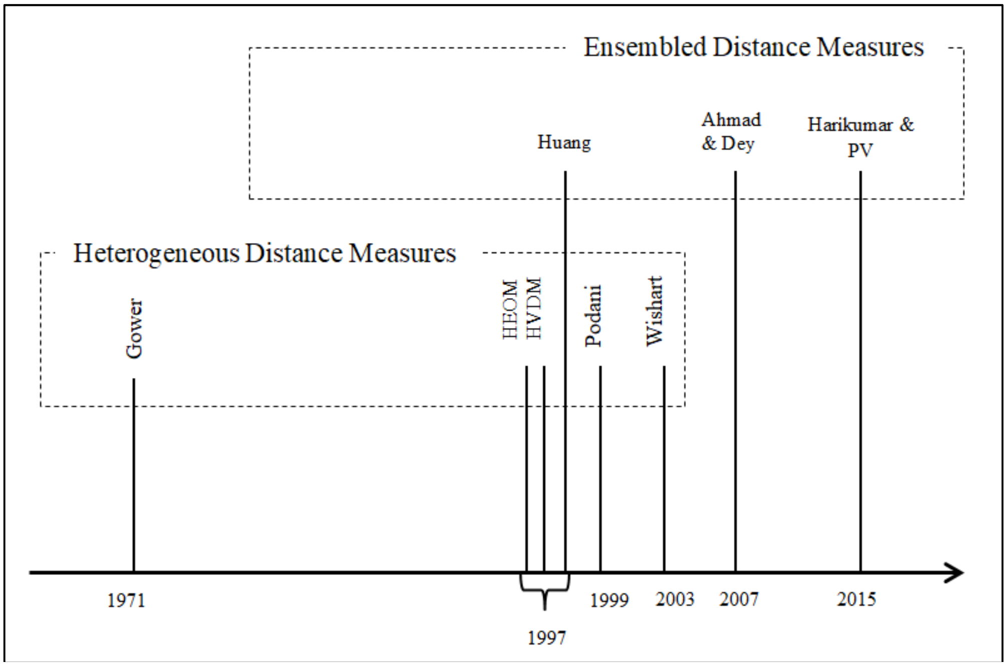

Fig 1: Chronological Table of Mixed Variables Distance Measures by Year

The two groups defined by us are clearly indicated in the figure above, and next begins the comprehensive discussion of these mixed variables distance measures starting with the earliest one.

Gower's coefficient

The very first distance measure for calculating the distance between two observations, which had continuous and categorical variables measured simultaneously, was proposed by Gower in 1971 [8]. Gower (1971) [8] defined a general coefficient which measures the similarity between two units, and this coefficient includes several existing ones as special cases, hence can be used under different circumstances.

The two individuals *i* and *j* can be compared on a variable *k* and assigned a score s_{ijk}. The similarity between *i* and *j* is defined as the weighted average score taken over all possible comparisons:

$$S_{ij} = \frac{\sum_{k=1}^{p} s_{ijk}\delta_{ijk}}{\sum_{k=1}^{p} \delta_{ijk}}$$

δ_{ijk} represents the possibility of making comparisons, that is δ_{ijk}=1 when variable k can be compared for *i* and *j*, which means no missing value for both. Sometimes no comparison is possible because of missing observation or in case of dichotomous variable where the agreement (0, 0) is considered non informative, in such cases δ_{ijk} is 0.

Compliment of S_{ij} is the distance measure, $d_{ij} = 1 - S_{ij}$

The scores s_{ijk} are assigned as:

a) Binary variables: If + is presence of character and – is its absence, then validity and score assigned to each combination is given as

	Values of character *k*
Individual *i* *j*	+ + - - + - + -
s_{ijk} δ_{ijk}	1 0 0 0 1 1 1 0

s_{ijk}=0 when i and j are considered different, s_{ijk} is unity when they have some degree of similarity.

b) Categorical variables: value are,

$s_{ijk} = 1$, if the two individuals *i* and *j* agree in *kth* character

$s_{ijk} = 0$, if they differ

c) Continuous variables: for continuous variables with values x_1, x_2,…….,x_n of character *k* for the total sample of *n* individuals,

$$s_{ijk} = 1 - \frac{|x_i - x_j|}{R_k}$$

$R_k$is the range of character *k* and may be the total range in the population, when x_i=x_j then s_{ijk}=1 and $s_{ijk}$is minimum when x_i and x_j are at opposite ends of their range.

Properties:

- S_{ij}ranges between 0 and 1, a value of 1 means that two individuals differ in no character and 0 means they differ maximally in all their variables.
- When δ_{ijk}=0 for all characters, S_{ij} is undefined and when all comparisons are possible $\sum_{k=1}^{v} \delta_{ijk} = p$, the total number of variables
- When there are no missing values, S_{ij} is positive semi-definite.

Gower coefficient is suitable for including in computer programs because it can cope up with different data types without any reprogramming. It has been found to be flexible to handle nearly all forms of character coding so far encountered, and unlike many coefficients does not require any recoding for quantitative characters. The important property of this similarity coefficient is the positive semi definite property, which allows numerical methods to be used with confidence that operate only on positive semi definite matrices provided there are no missing values. Missing values cause the similarity matrix to lose its positive semi definite property (Gower, 1971) [8].

Gower also suggested incorporating weights in the similarity coefficient, but deciding for weights is more difficult. According to him most simple weighting gives a constant weight to each character. But differences in characters may be considered more important than agreement, so weight for character k should be a function of character values x_{ik} and x_{jk} for individual i and j, and this functional form can be different for different characters. Many authors suggested different weights to the different variables for Gower's distance (Chae *et al*., 2006) [4].

Though Gower's work was directed towards taxonomists but it has impacted a much larger audience of various fields. His general coefficient has been used in different fields like medicine, genetics etc. (Cuadras, 1992b; Cuadras *et al.,* 1997; Mohammadi and Prasanna, 2003) [5, 6, 14].

Huang's distance

Huang (1997) [11] defined a distance measure for mixed variable data by combining the square Euclidean distance for numeric variables and simple matching distance for categorical variables.

$$d_{ij} = d_{ij}^{N} + \gamma d_{ij}^{C}$$

whered_{ij}^{N} is the distance between numeric variables,d_{ij}^{C} is the distance between categorical variables, and γ is the weight for categorical variables.

$$d_{ij}^{N} = \sum_{k=1}^{P_n} \left(x_{ik} - x_{jk}\right)^2$$

$$d_{ij}^{C} = \sum_{k=1}^{P_n} \delta_c(x_{ik}; x_{jk})$$

hereP_n are the number of numeric and P_n are the number of categorical variables,

$\delta_c(x_{ik}; x_{jk})$is the simple matching distance between object *i* and *j* in the categorical variable *k* and given as,

$$\delta_c(x_{ik}; x_{jk}) = \begin{matrix} 0, when\ x_{ik} = x_{jk} \\ 1, when\ x_{ik} \neq x_{jk} \end{matrix}$$

The Huang (1997) [11] distance between objects *i* and *j* is calculated by

$$d_{ij} = \sum_{k=1}^{P_n} (x_{ik} - x_{jk})^2 + \gamma \sum_{k=1}^{P_c} \delta_c(x_{ik}; x_{jk})$$

*γ*is the weight for categorical attributes, and is given as

$$\gamma = \frac{\sum_{k=1}^{P_n} s_k^2}{P_n}$$

It is introduced to avoid favoring either type of attributes. The choice of γ is dependent on the distributions of numerical attributes and generally it is taken as proportional to the average standard deviation of numeric attributes.

HEOM and HVDM

Stanfill and Waltz in 1986 introduced the distance measure for nominal variables called asoverlap distance which is the simplest measure of dissimilarity between two objects and it is simply the number of variables that are different between two objects, and it is given as

$$Overlap(x_i, x_j) = \begin{cases} 0 \; if \; x_i = x_j \\ 1 \; otherwise \end{cases}$$

$$\boldsymbol{d_{ij} = \sum_{k=1}^{p} (x_{ik} \neq x_{jk})}$$

Although this measure was simple but it was poor metric because it assigns an equal weight to all the variables and classes. Then, Stanfill and Waltz took a statistical approach to this problem by defining a new distance measure called Value Difference Metric (VDM).

$$\boldsymbol{d_{ij} = \sum_{k=1}^{p} \sum_{c=1}^{C} \left| \frac{N_{k,x_i,c}}{N_{k,x_i}} - \frac{N_{k,x_j,c}}{N_{k,x_j}} \right|}$$

where, $N_{k,x_i,c}$ is the number of objects that had value x_i for the variable *k* and an output class *c*

$N_{k,x_i}$is the number of objects that had value x_i for the variable *k*

$N_{k,x_j,c}$is the number of objects that had value x_j for the variable *k* and an output class *c*

$N_{k,x_j}$is the number of objects that had value x_j for the variable *k*

Value difference metric statistically determine the distance of two objects based on the proportion of the number of times their particular attributes are in the same class. Overlap distance measure and value difference metric handle only categorical variables and fail for continuous variables, because in continuous data there are very few overlaps.

Wilson and Martinez in 1997 [21] extended the overlap and VDM measure for the situations where the data ismix of categorical and continuous variables. They extended the overlap measure with the Heterogeneous Euclidean Overlap Metric (HEOM), which is defined as:

$$HEOM = \sum_{k=1}^{p} d_{ij}^2$$

$$d_{ij} = \sum_{k=1}^{p} \begin{cases} overlap(x_i, x_j) if \; the \; attribute \; is \; categorical \\ normdiff(x_i, x_j) if \; the \; attribute \; is \; continuous \end{cases}$$

where, $overlap(x_i, x_j)$ is same as defined above, and

$$normdiff(x_i, x_j) = \frac{|x_i - x_j|}{range_k}$$

The contribution of continuous variable is defined by the*normdiff* function, $range_k$ denotes the range of values of the *kth* variable. To mix continuous and categorical attributes, the continuous variables are normalized so that they do not have more or less weight than categorical attributes, as no categorical variable in overlap measure could contribute more than one to the distance.

Wilson and Martinez (1996) [20] discussed several possible ways to extend VDM to the continuous variables. First was to discretize the continuous variables, but by treating continuous variables as categorical a lot of information is lost. Then in 1997 they came up with another alternative to VDM which is called as Heterogeneous Value Difference Metric (HVDM) and is defined as:

$$HVDM(x_i, x_j) = \sqrt{\sum_{k=1}^{p} d_{ij}^2}$$

$$d_{ij} = \begin{cases} 1 \; if \; x_i \; or \; x_j \; is \; unknown \\ normalized\_vdm(x_i, x_j) if \; the \; attribute \; is \; categorical \\ normalized\_diff(x_i, x_j) if \; the \; attribute \; is \; continuous \end{cases}$$

where, $normalized\_diff(x_i, x_j) = \frac{|x_{ik} - x_{jk}|}{4\sigma_k}$, and σ_k is the standard deviation of continuous variable *k*. whereas several possibilities for *normalized_vdm* has been studied by the authors:

$$normalized\_vdm1(x_i, x_j) = \sum_{c=1}^{C} \left| \frac{N_{k,x_i,c}}{N_{k,x_i}} - \frac{N_{k,x_j,c}}{N_{k,x_j}} \right|$$

$$normalized\_vdm2(x_i, x_j) = \sqrt{\sum_{c=1}^{C} \left| \frac{N_{k,x_i,c}}{N_{k,x_i}} - \frac{N_{k,x_j,c}}{N_{k,x_j}} \right|^2}$$

Wilson & Martinez (1997) [21] tried HVDM over 15 different datasets and found that *normalized_vdm2* generalize the best. They compared it with HEOM and Euclidean, and HVDM was found to be superior.

Podani's distance

The Gower's general coefficient does not incorporate ordinal variables, which is a serious shortcoming if the mixed data sets have ordinal type variables. So as a solution, Podani in 1999 [15] extended Gower's general coefficient of similarity to ordinal characters. For ordinal variables$\delta_{ijk}$is the same as above and all $x_{ik}$are replaced by their ranks r_{ik}determined over all objects, and then

$$s_{ijk} = 1 - \frac{|r_{ik} - r_{jk}|}{max\{r_k\} - min\{r_k\}}$$

The idea here is to involve differences in ranks for two items within the same rank order, which is somewhat analogous to taking differences between the ranks for the same item in two orders, as in Spearman's rank correlation. Standardization by the range of ranks for each variable ensures comparability with the other variable types.

If ties appear, then correction terms are added to both the denominator and the numerator, given as

$$s_{ijk} = 1 - \frac{|r_{ik} - r_{jk}| - \frac{(T_{ik} - 1)}{2} - \frac{(T_{jk} - 1)}{2}}{max\{r_k\} - min\{r_k\} - \frac{(T_{k,max} - 1)}{2} - \frac{(T_{k,min} - 1)}{2}}$$

T_{ik} is the number of objects which have the same rank score for variable *k* as object *i*(including *i* itself), T_{jk} is the number

of objects which have the same rank score for variable k as object j(including j itself),
r_{ik}andr_{jk} are the maximum and minimum ranks for variable k, respectively,
$T_{k,max}$is the number of objects with the maximum rank, and $T_{k,min}$ is the number of objects with the minimum rank.
A distance coefficient alternative to Gower's index was given by Podani (2000) [16].

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \delta_{ijk} \left(\frac{x_{ik}-x_{jk}}{w_{ijk}}\right)^2}$$

whereδ_{ijk}= 0 if comparison of objects i and j for variable k is invalid for lack of data, otherwise δ_{ijk}= 1.
For different variables w_{ijk} is given as:
a) For binary variables: w_{ijk}= 1.
b) for nominal variables: $w_{ijk} = x_{ik} - x_{jk}$if $x_{ik} \neq x_{jk}$
w_{ijk}= 1 if $x_{ik} = x_{jk}$
c) for continuous variables: w_{ijk}= $max\{x_{ik}\} - min\{x_{jk}\}$
d) for ordinal variables: w_{ijk}= $max\{r_k\} - min\{r_k\}$

Wishart's distance

Wishart in 2003 proposed a distance measure which was similar to Gower's measure but with slight modification. He used variance of the continuous variable in the score part such that the distance is given as:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \delta_{ijk} \left(\frac{x_{ik}-x_{jk}}{w_{ijk}}\right)^2}$$

whereδ_{ijk}= 0 if comparison of objects i and j for variable k is not possible, otherwise δ_{ijk}= 1.
For different variables w_{ijk} is given as:

a. For binary/categorical variables: w_{ijk}= 1 if $x_{ik} = x_{jk}$, 0 otherwise
b. For continuous variables: w_{ijk}= s_k when k is numerical variable

Ahmad & dey's distance

Ahmad and Dey (2007) [1] proposed a distance measure which works well with mixed data, they somewhat modified the Huang's (1997) [11] distance measure. According to them most of the distance measures do not consider the distribution of values in the data set while computing the distance between a two categorical variables, which is naturally captured in case of continuous variables. They also illustrated that distance between values cannot be considered strictly binary because the values which co-occur together in same group should be more similar to each other than they were to the values occuring in different groups. The distance function of Ahmad and Dey used squared Euclidean distance between data objects for continuous variables and the distance between two categorical values is computed as function of their overall distribution and co-occurrence with other variables.

$$d_{ij} = \sum_{k=1}^{P_n}(x_{ik} - x_{jk})^2 + \sum_{k=1}^{P_c} \delta_c(x_{ik}; x_{jk})$$

where$\delta_c(x_{ik}; x_{jk})$ is the co-occurrence distance between categorical values.

Huang's measure used a binary valued distance for categorical variables and all categorical variables were weighted by a user defined parameter which controls the contribution of categorical variables to the distance function. But in Ahmad and Dey's measure the contribution of a categorical variable is inherent in the distance measure itself and is a function of co-occurrence of values. Thus, weighing values were extracted from the variable value distributions within the data. Also, the binary and categorical variables are not considered separately and the co-occurrence distance is based on both of these variables that are binary and categorical variables.

Harikumar and PV's distance

Harikumar and PV (2015) [9] proposed a generalized distance function for mixed data variables in the form of triple terms, which consists of three different distance measures for numeric, categorical and binary data types. According to them a small change in the dataset may change the results drastically if Euclidean distance is used. So, they used Manhattan distance for distance calculation because it is more flexible, robust and resistant to the outliers (Hopcroft & Kannan, 2013) [10]. Hamming distance was used for binary variables, and for categorical variables they used co-occurrence distance as defined by Ahmad and Dey (2007) [1]. Categorical variables were treated separately from binary variables due to the variations in their probability distributions unlike Ahmad and Dey who used same measure for binary and categorical variables. Another reason for treating the binary and categorical variables separately was that if binary variables were treated separately then the complex computations of calculating the probability of each variable value could be avoided.
Proposed generalized distance function in the form of triplet is given as:

$$d_{ij} = \sum_{k=1}^{P_n}|x_{ik} - x_{jk}| + \sum_{k=1}^{P_c} \delta_c(x_{ik}; x_{jk}) + \sum_{k=1}^{P_b} \delta_b(x_{ik}; x_{jk})$$

Here $\delta_b(x_{ik}; x_{jk}) = 0$ for $x_{ik} = x_{jk}$ and 1 otherwise by using Hamming distance.
So, the measure has three components, one for handling numeric variables, second for handling categorical variables and third for handling binary variables. For each component, lower value of distance indicates higher similarity.
Both Ahmad & Dey (2007) [1] and Harikumar & PV (2015) [9] used normalization scheme given by Witten & Frank (2000) [23] for numeric variables, and normalized value was given as:

$$d_{ij} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Packages in R for distance calculation

The distance measures which are discussed in this paper can be easily calculated in R software using some packages. The Gower's distance can be calculated using "gower" package in R, while the package "kmed" can be used to find the Huang's, Podani's, Wishart's, Ahmad & Dey's and Harikumar & PV's distance. The "UBL" package can be used to find the HEOM and HVDM measures.
Among the available distance measures there is no measure which can unanimously be considered as winner and this area needs more research. Hence no single distance measure is always superior or inferior because a distance measure that works well for one problem may be not good for other problem (Boriah *et al*., 2008; Veldon *et al.,* 2018) [2, 19].

Conclusions

This paper summarizes and provides comprehensive detail of various distance measures that can deal with mixed type of data including how the existing distance measures were modified to deal with mixed variables. Since such type of real world data occurs commonly in various fields, it is important to consider these distance measures. It will give an insight to the readers to take an informed decision in identifying the available options of distance measure and then selecting the appropriate measure for their specific problem.

References

1. Ahmad A, Dey L. A K-mean clustering algorithm for mixed numeric and categorical data. Data and Knowledge Engineering. 2007; 63:503-527.
2. Boriah S, Chandola V, Kumar V. Similarity measures for categorical data: A comparative evaluation. Conference: Proceedings of the SIAM International Conference on Data Mining, SDM, Atlanta, Georgia, USA, 2008.
3. Cha SH. Comprehensive survey on Distance/Similarity measures between probability density functions. International Journal of Mathematical Models and Methods in Applied Sciences. 2007; 1(4):300-307.
4. Chae SS, Kim JM, Yang WY. Cluster analysis with balancing weight on mixed-type data. The Korean Communications in Statistics. 2006; 13(3):719-732.
5. Cuadras CM. Some examples of distance based discrimination. Biometrical Letters. 1992b; 29:3-20.
6. Cuadras CM, Fortiana J, Oliva F. The proximity of an individual to a population with applications in discriminant analysis. Journal of Classification. 1997a; 14:117-136.
7. Gordan AD. Classification. London, 1981.
8. Gower JC. A general coefficient of similarity and some of its properties. Biometrics. 1971; 27:857-871.
9. Harikumar S, PV S. K-medoid clustering for heterogeneous data sets. Procedia Computer Science. 2015; 70:226-237.
10. Hopcroft J, Kannan R. Foundations of Data Science, 2013.
11. Huang Z. Clustering large data sets with mixed numeric and categorical values. The First Pacific Asia Conference on Knowledge Discovery and Data Mining, 1997, 21-34.
12. Mahalanobis PC. On the generalized distance in statistics. Proceedings of the Indian National Science Academy. 1936; 2:49-55.
13. Manly BFJ. Multivariate Statistical Methods: A Primer, 2nd edition. New York: Chapman & Hall, 1994.
14. Mohammadi SA, Prasanna BM. Analysis of genetic diversity in crop plants: Salient statistical tools and considerations. Crop Science. 2003; 43(4):1235-1248.
15. Podani J. Extending Gower's general coefficient of similarity to ordinal characters. Taxonomy. 1999; 48:331-340.
16. Podani J. Introduction to the exploration of multivariate biological data. Backhuys Publishers, 2000.
17. Sokal RR, Michener CD. A statistical methods for evaluating relationships. University of Kansas Science Bullentin. 1958; 38:1409-1448.
18. Stanfill C, Waltz D. Toward memory-based reasoning. Communications of the ACM. 1986; 29:1213-1228.
19. Velden MV, D'enza AI, Markos A. Distance-based clustering of mixed data, Computational Statistics. 2018; 11(3):101-126.
20. Wilson DR, Martinez TR. Heterogeneous radial basis functions. In Proceedings of the International Conference on Neural Networks. 1996; 2:1263-1267.
21. Wilson DR, Martinez TR. Improved heterogeneous distance functions. Journal of Artificial Intelligence Research. 1997; 6:1-34.
22. Wishart D. K-means clustering with outlier detection, mixed variables and missing values. In: Exploratory Data Analysis in Empirical Research: Studies in Classification, Data Analysis and Knowledge Organization, Springer, Berlin Heidelberg, 2003, 216-226.
23. Witten IH, Frank E. Data Mining Practical Machine Learning Tools and Techniques with Java Implementation. Morgan Kaufmann Publishers, San Francisco, CA, 2000.